

# Peripheral neuropathy outcome measures standardisation (PeriNomS) study part 2 : getting consensus

Citation for published version (APA):

Vanhoutte, E. K. (2015). *Peripheral neuropathy outcome measures standardisation (PeriNomS) study part 2 : getting consensus*. [Doctoral Thesis, Maastricht University]. Datawyse / Universitaire Pers Maastricht. <https://doi.org/10.26481/dis.20150313ev>

## Document status and date:

Published: 01/01/2015

## DOI:

[10.26481/dis.20150313ev](https://doi.org/10.26481/dis.20150313ev)

## Document Version:

Publisher's PDF, also known as Version of record

## Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

## General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

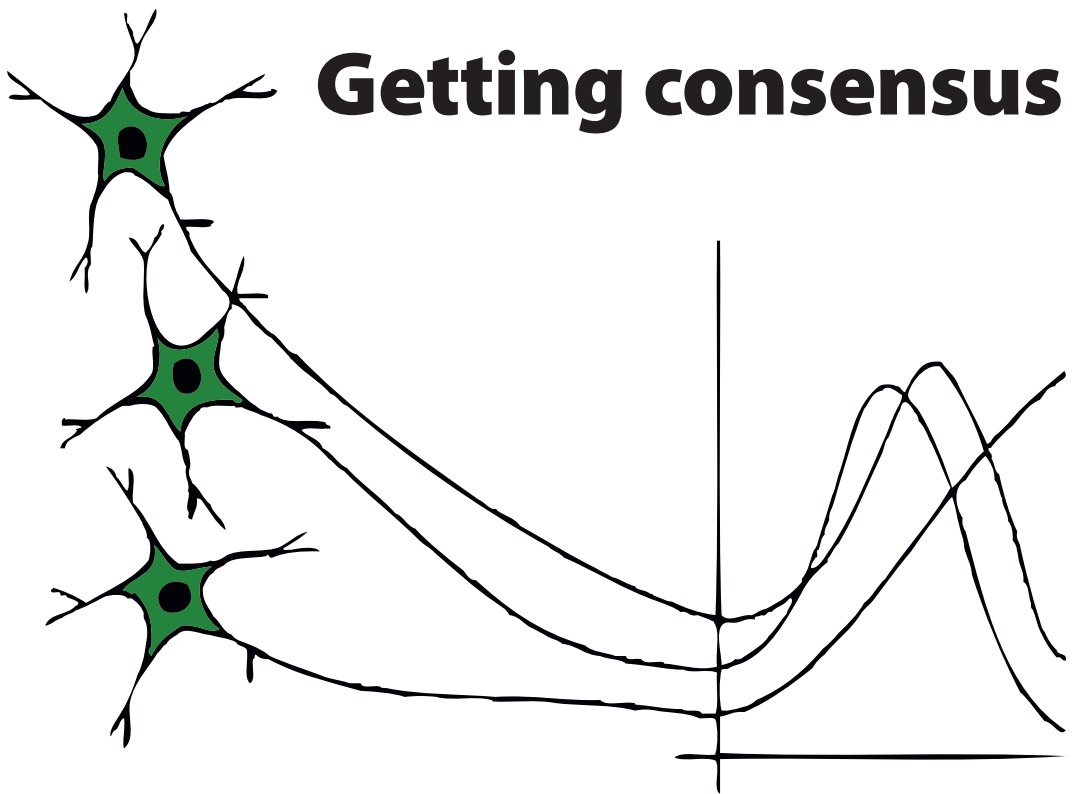
## Take down policy

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

Peripheral Neuropathy outcome measures  
Standardisation (PeriNomS) study part 2:



Els K. Vanhoutte

The studies described in this thesis were conducted in collaboration with 26 international neuromuscular centres as part of the PeriNomS study and coordinated from the Department of Neurology, Maastricht University Medical Centre, Maastricht, the Netherlands. This research project was financially supported (grants) by GBS/CIDP International Foundation, USA and Talents programm of Talecris industry, USA, now part of Grifols. Financial support for printing of this thesis was kindly provided by CSL Behring BV and Lamepro BV.

© E. K. Vanhoutte, 2015

ISBN 978 94 6159 406 8

**Lay-out & cover** Jos Hendrix

**Printing** Universitaire Pers Maastricht

**Peripheral Neuropathy outcome measures  
Standardisation (PeriNomS) study part 2:  
Getting consensus**

**Proefschrift**

ter verkrijging van de graad van doctor aan de  
Universiteit Maastricht  
op gezag van de Rector Magnificus  
Prof. dr. L.L.G. Soete  
volgens besluit van het College van Decanen,  
in het openbaar te verdedigen

op vrijdag 13 maart 2015 om 16.00 uur

door

**Els Karla Vanhoutte**



**Promotiecommissie****Promotor:**

Prof. dr. C.G. Faber

Prof. dr. P.A. van Doorn

**Copromotor:**

Dr. I.S.J. Merkies

**Beoordelingscommissie:**

Prof. dr. C.E.M. de Die-Smulders, voorzitter

Prof. dr. M. De Baets

Prof. dr. J.L. Thonnard (Institut de Neuroscience, Brussel, België)

Prof. dr. M. De Visser (Academisch Medisch Centrum, Amsterdam)

Prof. dr. J.S.H. Vles

Blowing the candle, to go....  
May the light of your persuasiveness and strength  
Captured by the flickering of this candle,  
Illuminate the dark site of these illnesses  
May your encouraging thoughts and belief  
Reinforce those that will unfortunately follow  
May your Pain, Fatigue as well as Other discomfort you experience  
Make us realize the commitment we as physicians have:  
To follow the light, from candle to go...  
To serve you and others to blow *the candle of discomfort*  
To aim for your smile as the best quality metric

Dr. I.S.J. Merkies



# CONTENTS

Abbreviations	8
<b>Part I      Introduction</b>	<b>11</b>
Chapter 1. Introduction of the Peripheral Neuropathy Outcome Measures Standardization (PeriNomS) Study and outline of the thesis	13
Chapter 2. Rasch-ionale for neurologists	31
Chapter 3. Defining a responder: its theoretical background	51
<b>Part II      Impairment studies</b>	<b>57</b>
Chapter 4. Improving Fatigue assessment in immune-mediated neuropathies: The modified Rasch-built Fatigue Severity Scale	59
Chapter 5. Vigorimeter Grip strength in CIPD: a responsive tool that rapidly measures the effect of IVIG- the ICE study	75
Chapter 6. Modifying the Medical Research Council grading system through Rasch analysis	87
Chapter 7. Comparing the NIS vs MRC and INCAT sensory scale through Rasch analyses	111
<b>Part III      Activity and participation studies</b>	<b>129</b>
Chapter 8. Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies	131
Chapter 9. Rasch-built overall disability scale for multifocal motor neuropathy (MMN-RODS)	147
Chapter 10. Changing outcome in inflammatory neuropathies: Rasch-comparative responsiveness	163
Chapter 11. Defining a responder through Rasch analyses	183
<b>Part IV      Summary and future perspectives</b>	<b>197</b>
Chapter 12. 196th ENMC International Workshop: 'Outcome Measures in inflammatory peripheral Neuropathies	299
Chapter 13. Summary, general discussion, and future perspectives	217
Chapter 14. Nederlandse samenvatting	229
Dankwoord	237
Curriculum vitae	241
List of publications	242
Valorisation - addendum	245



## Abbreviations

AIDP	Acute inflammatory demyelination polyradiculoneuropathy
ALDS	AMC Linear Disability Score
ANOVA	Analysis of variance
AUC	Area under the curve
Bn	Ability of persons
CIDP	Chronic Inflammatory Demyelinating Polyradiculoneuropathy
ClinJSc	Clinical Judgment Score
COPM	Canadian Occupational Performance Measure
CTT	Classical Test Theory
Di	Difficulty of items
DIF	Differential Item Functioning
EMG	Electromyography
ENMC	European neuromuscular centre
ES	Effect Size
EuroQoL	Euro Quality of Life scale
EuroQoL-5D	EuroQoL group-5D scale
FSS	9-item fatigue severity scale
GBS	Guillain-Barré syndrome
HRQoL	Health-related quality of life
ICC	Item Characteristic Curve
ICC	Intraclass correlation coefficient
ICE trial	IGIV-C CIDP efficacy trial
ICF	International Classification of Functioning, Disability and Health
IENF	Intraepidermal nerve fibre
IGIV-C	Intravenous human immune globulin, 10% caprylate/chromatography purified
INCAT	Inflammatory Neuropathy Cause and Treatment
I-RODS	Inflammatory-RODS
IRT	Item Response Theory
ISS	INCAT sensory sumscore
IVlg	Intravenous immunoglobulins
logit	log-odds unit
MAG	Myelin-associated-glycoprotein
MCID	Minimum clinically important differences
MCID-SE	Minimum clinically important differences using the obtained individual standard error
MGUS	Monoclonal gammopathy of undetermined significance
MGUSP	MGUS related polyneuropathy
mISS	Modified INCAT sensory sumscore
MMN	Multifocal Motor Neuropathy

MRC	Medical Research Council
MRC-ss	Medical Research Council sum-score
NCS	Nerve conduction studies
NHP	Nottingham health profile
NIS	Neuropathy Impairment Scale
NISm	Motor subset of NIS
NISs	Sensory subset of NIS
ODSS	Overall disability sumscore
ONLS	Overall neuropathy limitation scale
PCA	Principal component analysis
PeriNomS	Peripheral Neuropathy Outcome Measures Standardization
PGIC	5-points patient global impression of change
PI-NRS	Pain intensity numerical scale
PNS	Peripheral Nerve Society
PPCM	Personal patient-centred measures
PSI	Person Separation Index
QoL	Quality of life
RMC trial	Randomized Methotrexate CIDP trial
R-ODS	Rasch-built overall disability scale
RT	Rasch-Transformed
RT-mISS	Rasch-transformed modified INCAT sensory
RT-MRC	Rasch-transformed MRC sum-score
RUMM	Rasch unidimensional measurement model
SD	Standard deviation
SE	Standard error
SES	Self-evaluation scale
SF-36	Short form 36-item health survey
SFFS	Short-form fatigue scale
SFN	Small fibre neuropathy
SigChange	Significant change
SIP	Sickness impact profile
SRM	Standardized response mean
VAS	Visual analogue scale
VSN	Vereniging Spierziekten Nederland
WHO	World health organisation
WHOQoLbref	Short form of the WHO Quality of Life scale



# **Part I**

## **Introduction**



**Introduction of the  
Peripheral Neuropathy  
Outcome Measures  
Standardization (PeriNomS)  
study and outline  
of the thesis**

## Introduction

Peripheral neuropathies may cause a wide range and severity of deficits at the body level (impairments, e.g. weakness and sensory deficit, autonomic complaints), which may lead to problems in daily life (e.g., walking, washing) and social functioning (e.g. work activities) with a reduction in quality of life expectations.<sup>1-3</sup> These consequences have been the focus of numerous trials in peripheral neuropathies, aiming to ameliorate symptoms, functional deficits, or even find a cure for the disease. The choice of a proper outcome measure representing the level of interest to be examined is one of the most fundamental steps that should be taken in the design of clinical studies in peripheral neuropathies. Choosing a proper outcome is not only dependent on the proposed research purposes, but also, and perhaps more importantly, on the fulfilment of the clinimetric needs by the scale of interest.<sup>4,5</sup> With an increasing demand for accuracy, all outcome measures should be rigorously examined to determine their scientific soundness before being generally used. Physicians are generally well informed on the basic clinimetrics requirements of outcome measures such as being simple, communicable, valid, reliable and responsive. These requirements have been considered for years as the most important bricks to build upon a solid clinimetric foundation.<sup>5-7</sup> However, additional modern clinimetric essentials in the design and evaluation of outcome measures are important. The Rasch method has been gaining interest.<sup>8,9</sup>

The Peripheral Neuropathy outcome measures Standardization (PeriNomS) study has extended the aims accomplished by the Inflammatory Neuropathy Cause And Treatment (INCAT) study group. The INCAT group strived for standardisation of outcome assessment in patients with autoimmune peripheral polyneuropathies. During a first workshop in 2004, coordinated by the European Neuromuscular Centre (ENMC), the INCAT group succeeded in addressing the principles underlying the choice of outcome measures in various forms of peripheral neuropathies, particularly in Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathies (CIDP), monoclonal gammopathy of undetermined significance polyneuropathies (MGUSP) and multifocal motor neuropathies (MMN), and general consensus was reached regarding which outcome measures should be examined in these disorders. In addition, caveats were highlighted suggesting the development of more modern clinimetrically-based and specific metrics for these illnesses.<sup>10</sup>

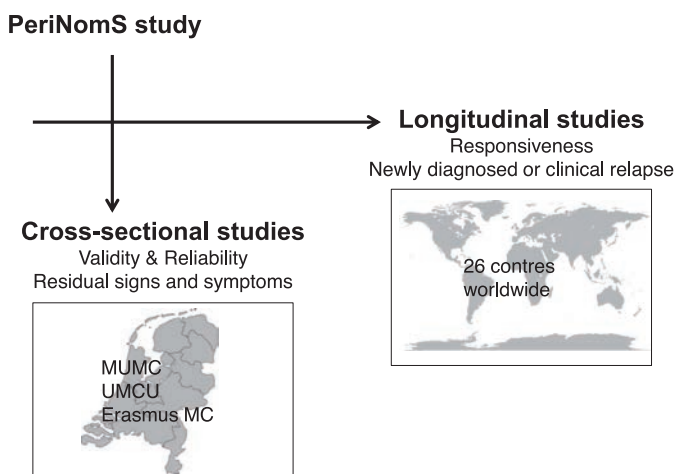
The PeriNomS study (2 years of preparation, 5 years of data collection; kick-off 2007, PNS, Utah, USA; dbase closed 31st December 2012) is an international collaborative effort of 26 neuromuscular centers with special interest in inflammatory neuropathies. Through comparative responsiveness studies, the most responsive measures were selected at the various levels of interest of assessing outcome. The study consisted of two parts: a cross-sectional part and a longitudinal part (Figure 1).

The aim of the PeriNomS study was to expand the clinimetric knowledge on outcome measures at selected levels of outcome in autoimmune polyneuropathies, particularly in GBS, CIDP, MMN and MGUSP. The goal was to present a minimum core set of outcome measures to be used in future follow-up studies and clinical trials in patients with an immune-mediated polyneuropathy, and to define the most responsive measures for each polyneuropathy form. This core set should be able to detect clinically important improvement or deterioration over time, which enables future clinical trials to be compared and this will improve the quality of future randomised controlled trials.

## Design

The PeriNomS study consisted of a cross-sectional (validity and reliability) and a longitudinal (responsiveness comparison) part.

**Figure 1.** Design of the PeriNomS study



### Cross-sectional studies

The cross-sectional part focused on examining validity and reliability in clinical stable patients with residual signs and symptoms. A total of 122 patients with an immune-mediated polyneuropathy (GBS: 30, CIDP: 30, MGUSP: 20, MMN: 22, inflammatory small fiber neuropathy (SFN): 20) and a stable clinical condition with residual symptoms and signs were examined at two different occasions at the outpatient clinics in the Netherlands. During the first visit two well-trained and independent observers performed their scores independently and consecutively, to obtain inter-observer



reliability scores. Within 2-4 weeks one of the investigators re-examined the patient (intra-observer values; equally distributed among the examiners) without having access to previous results. Patient-reported outcomes were assessed twice by patients in the same way, without having access to the previous scores.

### Longitudinal studies

These studies were performed worldwide (at the collaborating centers) and the data were centrally collected (Erasmus MC and Maastricht UMC, the Netherlands), and were used to determine comparative responsiveness scores between the ~equally valid and reliable pre-selected outcome measures that are generally being applied in immune-mediated polyneuropathies.<sup>1,6,10</sup>

A total of 163 newly diagnosed patients or patients with a clinical relapse were included (GBS: 55, CIDP: 59, MGUSP: 23 of which 19 were IgM anti-myelin associated glycoprotein positive, MMN: 26). During a one year follow-up patients with GBS of CIDP were examined 5 times: at onset, and after 1, 3, 6 and 12 months; patients with MMN and MGUSP were examined 3 times: at onset, and after 3 and 12 months. Basic characteristics of these patients were presented including treatment regimens if reported.

## Patients - Clinical description

### Guillain-Barré syndrome (GBS)

With an incidence of 1-2 per 100,000 persons a year, the Guillain-Barré syndrome (GBS) is the most common cause of acute flaccid paralysis.<sup>11,12</sup> The hallmark of GBS is a bilateral sub-acute flaccid paralysis of the limbs with decreased or absent deep tendon reflexes. The clinical presentation is highly variable, but generally consists of an acute, symmetrical ascending weakness of the limbs which can vary from mild to complete tetraplegia with respiratory failure. There are usually paresthesias, and numbness, and patients may have pain. After reaching a nadir in 2-4 weeks, recovery starts after a plateau phase of variable duration.<sup>13</sup>

Guillain-Barré syndrome is considered to be a spectrum of acute inflammatory neuropathic diseases. The acute inflammatory demyelination polyradiculoneuropathy (AIDP) is the most common subtype in Europa and North America.<sup>14,15</sup>

Intravenous Immunoglobulin and plasma exchange started within two weeks from onset may improve recovery. Despite this treatment many patients still have poor long-term outcome (morbidity and mortality)<sup>16</sup> Guillain-Barré syndrome is a monophasic illness and there are several factors associated with poor prognosis like older age, preceding diarrhea and severe disease.<sup>17</sup> GBS remains a severe illness since 25% of the patients require artificial ventilation and between 4 and 15% die.<sup>14,18</sup>

### **Chronic inflammatory demyelination polyradiculoneuropathy (CIDP)**

CIDP is characterized by symmetric muscle weakness, often with sensory disturbances and ataxia, and decreased or absent deep tendon reflexes. In most cases, it begins with paraesthesias and weakness in the distal limbs as well as difficulty walking. Usually, symptoms progress over > 8 weeks period. The disease course is steadily or stepwise progressive over at least 2 months, but can also be relapsing. In contrast with GBS, cranial nerves are rarely affected and respiratory or autonomic involvement is exceptional.<sup>19</sup> Diagnosis is based on a combination of clinical, electrodiagnostic, and laboratory features with the exclusion of disorders that might mimic CIDP.<sup>20</sup>

Depending on the diagnostic criteria used, the prevalence ranges from 2-5 per 100.000.<sup>21</sup> Therapy for disabling symptoms includes intravenous immunoglobulin, corticosteroids and plasma exchange.<sup>20</sup> Using one of these treatments, 50-80% of the patients have shown improvement.<sup>22</sup> Furthermore, long term treatment with immunoglobulins causes remission in a third of CIDP patients without the need of additional IVIg infusions, and more than half of the patients remained stable.<sup>23</sup>

### **Monoclonal gammopathy of unknown significance polyneuropathy (MGUSP)**

A paraprotein is an abnormal protein that is produced by bone marrow cells and can be divided into three classes: IgG, IgA and IgM.

IgM is produced by a clone of B-lymphocytes that are initially benign. Monoclonal gammopathies are a heterogeneous group of disorders with the presence of a paraprotein, ranging from the subclinical monoclonal gammopathy of unknown significance (MGUS) to malignancies such as multiple myeloma.<sup>24</sup> There is an association between IgM monoclonal gammopathy and neuropathy.<sup>24,25</sup> Monoclonal gammopathy of undermined significance polyneuropathy (MGUSP) usually occurs in older males, in the 6th to 9th decades of life.<sup>24</sup> The neuropathy is predominantly sensory with mild distal weakness with prolonged distal motor latencies, that can be accompanied by tremor and ataxia.<sup>20</sup> The neuropathy usually has an indolent progressive course. When severe disability or fast progression is present, immune-modulating treatment could be considered although scientific evidence is limited. There is a 1% annual risk that the MGUS transforms to malignancy, thus when fast progression is present, hematological follow-up is advised.<sup>20</sup> Patients might benefit from Rituximab, though recent papers did not show unequivocal therapeutic evidence for efficacy in this condition.<sup>26,27</sup> Older age at onset and demyelination are associated with an increased risk of more disability, whereas the presence of anti-MAG antibodies decreases this risk.<sup>28</sup>

### **Multifocal motor neuropathy (MMN)**

Multifocal motor neuropathy (MMN) is considered an immune-mediated disorder, although the initial and ultimate targets are unknown, it seems to be related to an immune attack on motor nerves.<sup>29</sup> MMN affects 1-2 persons per 100.000, is more frequent in man than women and has a mean age of onset around 40 years.<sup>30</sup>

MMN is characterized by slowly progressive, predominantly distal, asymmetric limb weakness and usually multiple partial motor nerve conduction blocks.<sup>29,31,32</sup> The onset of weakness is typically insidious with slow progression and often starts with wrist drop, grip weakness and foot drop.<sup>29,32,33</sup> Pain or sensory symptoms like numbness and tingling are generally absent.<sup>29</sup> Diagnostic criteria proposed by Asseltonk and associates consist of a combined clinical, laboratory, and electrophysiological characteristics for definite, probable, and possible MMN.<sup>32</sup>

MMN may respond to immunomodulatory treatments; spontaneous improvement has rarely been described.<sup>29,34,35</sup> Intravenous immunoglobulin has been reported to be efficacious in most patients, but long-term maintenance therapy does not prevent slowly progressive axonal degeneration.<sup>36-40</sup>

## Patients - Recruited

### Cross-sectional studies

For the cross-sectional studies, we included 122 patients with the following diseases: GBS (n=30), CIDP (n=30), MGUSP (n=20), MMN (n=22).<sup>41-44</sup> However, for the purposes of this thesis, the findings in the patients with SFN will not be addressed.

Only patients with a clinical stable condition will be included for these studies. A clinical stable condition is defined as:

- An unchanged clinical functionality as declared by the patient to the best of his/her knowledge over 2 months prior to the study and during the study (stable condition at second visit compared to entry).
- No clear objective changes at neurological examination by the researcher when compared with recorded findings over two month prior to study entry (if available) and during the study (unchanged neurological examination at second visit compared to entry).

The selected patients had residual symptoms or signs due to their illness that represents a broad range of activity limitations before admission to the current study.<sup>1</sup> Patients with CIDP or MMN receiving (interval) therapy were included as long as their clinical condition was stable. Patients with MGUSP with preferably serologically proven IgM anti-MAG+ antibodies were included in the study. Patients with a clinical diagnosis and a clinical course compatible with CIDP and an IgG MGUS were included as CIDP. Patients were excluded from participation if there were concomitant diseases like, e.g., diabetes, renal insufficiency, (prior) treatment with chemotherapy, alcohol abuse (more than 5 IU/day) that might interfere with general nervous system as well as physical functioning.

The patients were recruited from the Rotterdam, Maastricht, and Utrecht (inflammatory) neuropathy databank.

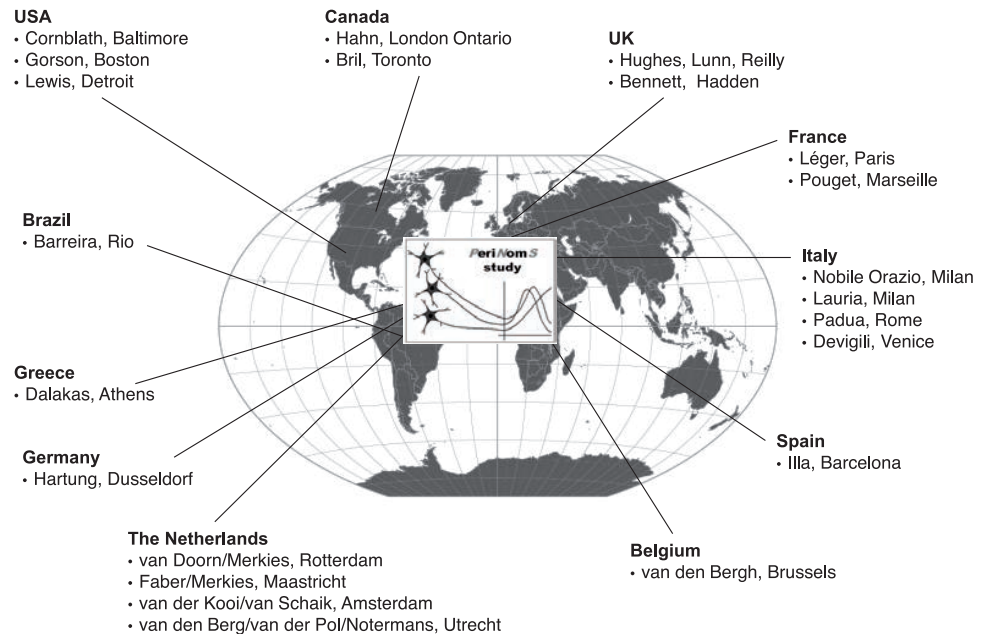
**Table 1.** Inclusion per country

Country	GBS	CIDP	MGUSP	MMN	Total
The Netherlands	25	20	8	9	62
USA	13	16	4	5	38
Italy	9	9	3	5	26
Belgium	2	2	5	1	10
France	1	4	1	3	9
UK	2	2	1	2	7
Canada	1	2	1	1	5
Brazil	2	2	0	0	4
Spain	0	2	0	0	2
Total	55	59	23	26	163

## Longitudinal studies

Worldwide, 26 centers with expertise on peripheral neuropathies included a total of 163 patients with GBS, CIDP, MGUSP, or MMN (Figure 2 and Table 1).

**Figure 2.** The Peripheral Neuropathy Outcome Measures standardization (PeriNomS) study group



### Inclusion / Exclusion criteria:

All patients were newly diagnosed, or were having a clinical relapse and all patients met the international criteria for their diagnosis.<sup>10,41,43,45</sup>

Patients with CIDP who demonstrated a clinical relapse, and were not using any medication for their polyneuropathy (e.g. intravenous immunoglobulins, steroids, immunosuppressive agent) for at least 2 months prior to inclusion, were eligible for the longitudinal part of the study. Patients with MMN having a new (multi)focal nerve lesion, without using any medication for their polyneuropathy (e.g. intravenous immunoglobulins) for at least 2 months prior to inclusion, were eligible as well. MGUSP patients were only included if they had serologically proven IgM and preferably also anti-MAG positive antibodies. MGUSP patients with an indication for therapy were

**Table 2.** Selected outcome measures in PeriNomS study

Outcome measure		Cross-sectional studies		Longitudinal studies		
		GBS/CIDP/MGUSP	MMN	GBS/CIDP	MGUSP	MMN
<b>Pathology</b>	IENF density	1x	-	2x	2x	-
<b>Impairment</b>	Jamar dynamometer	3x	3x	5x	3x	3x
	Vigorimeter	3x	3x	5x	3x	3x
	MRC sum-score	3x	3x	5x	3x	3x
	NIS motor subset	3x	3x	5x	3x	3x
	INCAT sensory sum-score	3x	-	5x	3x	-
	NIS sensory subset	3x	-	5x	3x	-
	EMG	1x	1x	2x	2x	2x
	11-point PI-NRS	2x	-	5x	3x	-
	VAS-pain	2x	-	5x	3x	-
<b>Activity and participation</b>	ODSS	3x	3x	5x	3x	3x
	ONLS	3x	3x	5x	3x	3x
	Inflammatory-RODS/MMN-RODS	2x	2x	5x	3x	3x
<b>Quality of life</b>	Vickrey's scale	2x	2x	5x	3x	3x
	SF-36	2x	2x	5x	3x	3x
	EuroQoL-5D	2x	2x	5x	3x	3x
	SIP	2x	2x	5x	3x	3x
	NHP	2x	2x	5x	3x	3x
	WHO-QoL bref	2x	2x	5x	3x	3x
	VAS-QoL	2x	2x	5x	3x	3x
	PPCM	-	-	5x	3x	3x
<b>Composite Levels</b>	Clinical judgment score	-	-	5x	3x	3x

**Legend to Table 2:** IENF: intra-epidermal nerve fiber, MRC: Medical Research Council, NIS: neuropathy impairment score, INCAT: Inflammatory Neuropathy Cause and Treatment group, EMG: electromyography, PI-NRS: pain intensity numerical scale, VAS: visual analogue scale, ODSS: overall disability sum-score, ONLS: overall neuropathy limitations scale, inflammatory-RODS: Rasch-built overall disability scale for GBS/CIDP/MGUSP, MMN-RODS: Rasch-built overall disability scale specifically for MMN, Vickrey's scale is a n=97 items peripheral neuropathy quality of life instrument, SF-36: short form 36-item health survey, EuroQoL-5D: EuroQoL group-5D scale, SIP: sickness impact profile, NHP: Nottingham health profile, WHO-QoL brief: short form of the WHO quality of life scale, QoL: quality of life, PPCM: personal patient-centered measures.

preferably included (on basis of clear progression of symptoms). However, IgM anti-MAG+ MGUSP patients not receiving therapy could also be included to study the natural course. Patients with a clinical diagnosis and a clinical course compatible with CIDP and IgG MGUS were included as CIDP.

Patients with GBS, CIDP, MGUSP, and MMN were excluded from participation if there were concomitant diseases like diabetes, renal insufficiency, (prior) treatment with chemotherapy, alcohol abuse (arbitrarily chosen as more than 5 IU/day) that might interfere with general nervous system as well as physical functioning. Patients with an IgM MGUSP without serologically proven IgM anti-MAG+ antibodies were excluded from the study.

Thus, excluding the patients with SFN, a total of 306 assessments were performed in the cross-sectional part. A total of 717 assessments were scheduled in the longitudinal part and 93% of the data were collected (in GBS/CIDP/MGUS: n=600 assessments, in MMN: n=68 assessments performed).

A comprehensive training was conducted (PNS congress 2011, Potomac, Washington, USA) for all participating centers, aiming to standardize the assessment procedures for all selected measures. A research manual was also provided to all participating centers explaining the various steps needed to increase reliability when assessing an outcome in these patients. The manual was constructed after having achieved consensus regarding aspects such as position of patient and physician in the manual assessment of muscle groups, location to stabilize the patient and counter-force location.

## **Selected outcome measures**

### **Selected outcome measures at the pathology level:**

- **Intra-epidermal nerve fiber (IENF) density**

Epidermal nerves are the distal terminals of small dorsal root ganglia neurons that pierce the dermal-epidermal basement membrane and penetrate the epidermis. A punch biopsy was performed following established procedures, 10 cm above the lateral malleolus in both legs after local anesthesia with 1% lidocaine.<sup>46</sup> Validated normal values were used.<sup>47</sup> Since determination of IENF density is not part of routine diagnostic procedure, it was only performed centers with expertise (the Netherlands and Italy) after obtaining written informed consent.

### **Selected outcome measures at the impairment level:**

- **Electromyography (EMG)**

Standardized electrophysiological measurements were performed in the right arm and leg using conventional techniques with surface electrode recordings after limb warming-up procedure.<sup>48,49</sup> Tests included assessment of motor nerves: median

(two-point stimulation: wrist and distal to elbow), ulnar (three-point stimulation: wrist, distal and proximal to elbow), peroneal (three-point stimulation: ankle, fibular head, popliteal fossa), and tibial (two-point stimulation: ankle, popliteal fossa) to determine the compound muscle action potential following distal and proximal stimulation, terminal latencies, conduction velocities, and F-response latencies. Sensory nerves (median (II), ulnar (V) and sural nerves) were stimulated antidromically to determine evoked sensory nerve action potential amplitudes, distal latencies and conduction velocities. Needle electromyography of first dorsal interosseous, vastus medialis and anterior tibial muscles were carried out to evaluate the presence of fibrillation potentials and positive sharp waves.

- **Medical Research Council sum-score (MRC-ss)<sup>50</sup>**

The MRC scale was used to score each muscle group and the scores are given in full numbers only, with 0=no movement, no contraction; 1=visible contraction without movement, 3=movement against gravity; 4= movement against resistance, but weaker than normal; 5=normal strength.

In GBS, CIDP and MGUSP, the MRC sum-score (MRCss) was composed by the summation of the following six muscle pairs: shoulder abduction, elbow flexion, wrist extension, hip flexion, knee extension and ankle dorsiflexion. This sum-score ranges from 0 (total paralysis) to 60 (maximal strength).

In MMN, a modified MRC sum-score covering fourteen predefined pairs of as proposed by the 131st ENMC International workshop was used.<sup>10</sup> The following muscle groups were chosen based on experts opinion: shoulder abductors, elbow extensors, and flexors, wrist extensors and flexors, finger extensors and flexors, thumb abductors, hypothenar abductors, hip flexors, knee extensors and flexors, ankle dorsiflexors and ankle plantar flexors. This sum-score ranges from 0 (total paralysis) to 140 (normal strength).

- **Neuropathy Impairment Scale (NIS)<sup>51-53</sup>**

The motor and sensory subsets of this scale were assessed according to earlier published procedures.<sup>51-53</sup> Sensation, consisting of touch pressure (range 0-8), pinprick (range 0-8), vibration (0-8) and joint position (range 0-8), was tested at the index finger and at the great toe on both sides. Touch pressure was assessed with long fiber cotton wool; pinprick with straight pins, vibration sensation with a 128 Hz tuning fork and joint position by moving the terminal phalanx of the index finger and great toe. Scoring was as follows: 0=normal, 1=decreased and 2= absent sensation. The NIS sensory ranges from 0 (no sensory deficit) to 32 (maximal sensory deficit).

Manual muscle strength was graded as 0=normal, 1=25% weak, 2=50% weak, 3=75% weak, 3.25=move against gravity, 3.5=movement, gravity eliminated, 3.75=muscle flicker, no movement, 4=paralysis. In GBS, CIDP and MGUSP, 12 muscles were tested (arm abductors, forearm flexors, wrist extensors, hip flexors, knee extensors, foot dorsal flexors at both sides) and the NIS-motor ranged from 0 (no motor deficit) to 48

(most severe motor deficit). In MMN 28 muscles were examined (at both sides: arm abductors, elbow extensors and flexors, wrist extensors and flexors, finger extensors and flexors, thumb abductors, hypothenar abductors, hip flexors, knee extensors and flexors, ankle dorsiflexors and ankle plantar flexors) and the NIS-motor ranged from 0 (no motor deficit) to 112 (most severe motor deficit). The examiner scored deficits by what he (she) considered to be normal considering test, anatomical site, age, gender, weight, and physical fitness.

- **Vigorimeter and Jamar dynamometer<sup>1,54-56</sup>**

With these two instruments grip strength was measured in both hands and the mean of three trials was used. The Jamar (Sammons Preston Rolyan, Bolingbrook, IL, USA) measures isometric strength and ranges from 0 (no grip strength) to 200 lbs/90 kg (dual scale). The pressure in the bulb of the Vigorimeter (Martin, Tuttlingen, Germany) is registered on a manometer via a rubber junction tube and ranges from 0 (no grip strength) to 156 (kilopascal).

All patients were asked to judge both the Vigorimeter and the Jamar regarding their preference for one of these instruments in assessing grip force when compared to each other (1 = great preference for Vigorimeter; 2 = little preference for Vigorimeter; 3 = no preference for any of the two instruments; 4 = little preference for Jamar; 5 = great preference for Jamar).

- **INCAT sensory sum-score (ISS), modified<sup>1,10,57</sup>**

Based on suggestions of the 2004 ENMC workshop, the ISS was expanded with light touch and joint position sense, since this would most probably make this scale more comprehensive and perhaps more responsive.<sup>1,10,57</sup> In addition, the scoring of the 2-point discrimination was also changed and based on the collected normative values.<sup>1,4</sup>

The modified ISS ranges from 0 (no sensory deficit) to 33 (most severe sensory deficit) and is composed by the summation of pinprick arm grade (range 0-4), vibration arm grade (range 0-4), light touch arm grade (range 0-4), joint position arm grade (grade 0-4), two-point discrimination grade (range 0-1), pinprick leg grade (range 0-4), vibration leg grade (range 0-4), light touch leg grade (range 0-4) and joint position leg grade (grade 0-4). The grades were defined as: normal (grade 0) or disturbed (grade 1) sense at the distal interphalangeal joint of the index finger or hallux; abnormal sense at the wrist or ankle (grade 2), at the elbow or knee (grade 3), at shoulder or hip joint (grade 4). Light touch and pin prick examinations were performed using the calibrated neuropen (10 grams monofilament for touch plus neurotips for pin examination).<sup>58,59</sup>

- **Visual Analogue pain Scale (VAS) and 11-point-pain intensity numerical scale (PI-NRS)<sup>60,61</sup>**

At each visit, all patients, except MMN patients, were instructed to complete the VAS and PI-NRS pain scales according to published procedures.<sup>60,61</sup>



### **Selected outcome measures at the activity and participation level**

- **Overall disability sum-score (ODSS) and Overall Neuropathy Limitation Scale (ONLS)<sup>1,62</sup>**

These activity limitation scales were completed according to the published procedures by each one of them and are composed by a summation of an arm disability scale (range 0-5) and a leg disability scale (range 0-7).<sup>1,62</sup> The examiner completed the scale after having interviewed the patient. If needed, the patient may be observed performing a task.

- **Rasch-built overall disability scale (R-ODS)<sup>63</sup>**

Based on caveats in the literature, two Rasch-built overall disability measures were constructed, one specifically for patients with GBS, CIDP, and MGUSP (inflammatory-RODS) and one specifically for patients with MMN (MMN-RODS). Their construction and clinimetric findings will be addressed in chapters 8 to 11.

### **Selected outcome measures at the quality of life level**

Except for the Vickrey's quality of life scale, there is no specific peripheral neuropathy quality of life measure.<sup>64</sup> In addition, there is also no consensus regarding which outcome measure at this level should be preferably assessed in these conditions.<sup>65</sup> Therefore, we have decided to collect data on most quality of life measures known, aiming to pool the data and to construct an inflammatory neuropathy specific Rasch-built measure. The following measures were assessed: Vickrey's scale, the medical outcome study short form 36 items health status (SF-36), Euro Quality of Life scale (EuroQoL), Sickness Impact Profile (SIP), Nottingham health profile (NHP), short form of the WHO Quality of Life scale (WHOQoLbref) are all health-related quality of life (HRQoL) measures.

The Vickrey's scale is a HRQoL measure specific for peripheral neuropathy.<sup>64</sup> The other scales concern generic quality of life measures.<sup>66-77</sup> All HRQoL measures were completed according to the published procedures by each one of them.<sup>64,66-77</sup>

### **Selected outcome measures to judge clinical state by the individual patient**

- **Clinical Judgment Score (ClinJSc)**

At each visit, patients were requested to judge whether their clinical condition strongly deteriorated (coded 1), slightly deteriorated (coded 2), remained stable (coded 3), slightly improved (coded 4), or strongly improved (coded 5) when compared with last visit (= defined as 'clinical-judgment-score'). At study entry, the patients reflected their clinical condition against their physical status within the two weeks before the start of the study.

- **Personal patient-centred measures (PPCM)**

These are based on the Canadian Occupational Performance Measure (COPM) concept of self-reported activities using a semi-structured interview undertaking

the following steps: the researcher helped the patient to identify problem areas in daily activities and social participation. A problem exists when a patient cannot do, doesn't do or isn't satisfied with how he is doing his occupation or a particular task. By focusing on the patients' own roles and environments we tried to identify areas that are relevant to them. After identifying problem areas, patients were asked to order the selected problems by their priorities (from most important to less important). Subsequently, the 5 most important problem areas to the patient were selected for follow-up. These items were scored on a 5 point rating scale (1 meaning "not able to do" or "not satisfied at all" to 5 meaning "able to do it extremely well" or "extremely satisfied") depending on the personal performance and satisfaction by the patient.<sup>78,79</sup>

## Results of the PeriNomS study

### Part I. Improving and standardising assessment of patients with immune-mediated neuropathies

Presented by dr. Sonja I. van Nes who defended her thesis in January 2011.<sup>80</sup> The main focus was on the cross-sectional part of the PeriNomS study. Various normative values were collected, particularly for the 2-point discrimination and the Jamar.<sup>4,56</sup> The 2-point discrimination normative values were subsequently incorporated as part of the modified INCAT sensory sum score in addition to adding the joint and touch sense to this scale.<sup>57</sup> Basic fundamental aspects between classical test theory ordinal-based measures versus modern clinimetrically constructed metrics (represented by Rasch methodology) were addressed, the advantages and drawbacks of each method being highlighted.<sup>63</sup> The Rasch methodology was used to construct the inflammatory Rasch-built overall disability scale (inflammatory-RODS) based on preliminary longitudinal data. The I-RODS demonstrating a better targeting of the ability levels of patients with GBS, CIDP, and MGUSP when compared to the widely used INCAT-ODSS disability ordinal-based measures.<sup>63,81,82</sup> The first comparative validity and reliability studies were presented for some of the selected (modified) existing and newly devised outcome measures. Two interval grip strength measures were compared, the Jamar dynamometer to the Vigorimeter in patients with GBS, CIDP, MGUSP and MMN. No difference in validity and reliability was seen between both tools. Also, a traditional comparison was made between the sensory subset of the neuropathy impairment score (NISs) and the modified INCAT sensory sumscore (mISS). No difference in validity and reliability was seen. The concept of minimum clinically important difference (MCID) and using it to define a responder was introduced.<sup>83</sup> Furthermore, the first comparative responsiveness preliminary results of the longitudinal part of the PeriNomS study were presented using the MCID concept.

## **Part II. Peripheral Neuropathy outcome measures Standardisation (PeriNomS) study part 2: getting consensus**

This thesis combines the cross-sectional and longitudinal data to describe extensively the final responsiveness comparative results between equally valid and reliable measures as the second step of the PeriNomS study. For each disease (GBS, CIDP, MGUSP, and MMN, separately) a final specific minimum core set of outcome measures was agreed upon between neuromuscular experts in the inflammatory neuropathy field. The disease specific core-set of outcome measures was reached through a consensus workshop, held under the auspice of the ENMC in 2013.<sup>84</sup>

This thesis addresses the following aspects:

- Introduction and background information
  - Rasch-ionale for neurologists (Chapter 2)
  - Defining a responder: its theoretical background (Chapter 3)
- Impairment studies
  - Improving Fatigue assessment in immune-mediated neuropathies: The modified Rasch-built Fatigue Severity Scale (Chapter 4)
  - Vigorimeter Grip strength in CIDP: a responsive tool that rapidly measures the effect of IVIG- the ICE study (Chapter 5)
  - Modifying the Medical Research Council grading system through Rasch analysis (Chapter 6)
  - Comparing the NIS vs MRC and INCAT sensory scale in inflammatory neuropathies through Rasch analyses (Chapter 7)
- Activity and participation studies
  - Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies (Chapter 8)
  - Rasch-built overall disability scale for multifocal motor neuropathy (RODS-MMN) (Chapter 9)
  - Changing outcome in inflammatory neuropathies: Rasch-comparative responsiveness (Chapter 10)
  - Defining a responder through Rasch analyses (Chapter 11)
- Summary and future perspectives
  - 196th ENMC International Workshop: 'Outcome Measures in inflammatory peripheral Neuropathies (Chapter 12)
  - Summary, general discussion, and future perspectives (Chapter 13)

## **Part III: Peripheral Neuropathy outcome measures Standardisation (PeriNomS) study part 3**

The remaining future needs will be addressed as part of the final third part (see future perspectives, chapter 13)

## References

- 1 Merckies, I. S. J. *Evaluation of scales and measurement instruments in immune-mediated polyneuropathies*. Thesis, Erasmus Medical Center, (2001).
- 2 Aaronson, N. K. Quality of life: what is it? How should it be measured? *Oncology (Williston Park)* **2**, 69-76, 64 (1988).
- 3 England, J. D., Gronseth, G. S., Franklin, G. *et al.* Distal symmetric polyneuropathy: a definition for clinical research: report of the American Academy of Neurology, the American Association of Electrodiagnostic Medicine, and the American Academy of Physical Medicine and Rehabilitation. *Neurology* **64**, 199-207 (2005).
- 4 van Nes, S. I., Faber, C. G., Hamers, R. M. *et al.* Revising two-point discrimination assessment in normal aging and in patients with polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **79**, 832-834 (2008).
- 5 Streiner, D. L. and Norman, G. R. *Health measurement scales: a practical guide to their development and use*. 2nd edn, (Oxford University Press, USA, 1998).
- 6 Feinstein, A. R. *Clinimetrics*. (Yale University Press, 1987).
- 7 Liang, M. H. Evaluating measurement responsiveness. *J Rheumatol* **22**, 1191-1192 (1995).
- 8 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 9 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1960).
- 10 Merckies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 11 Pritchard, J. and Hughes, R. A. Guillain-Barre syndrome. *Lancet* **363**, 2186-2188 (2004).
- 12 van der Maas, N. A., Kramer, M. A., Jacobs, B. C. *et al.* Guillain-Barre syndrome: background incidence rates in The Netherlands. *J Peripher Nerv Syst* **16**, 243-249 (2011).
- 13 Van der Meche, F. G., Van Doorn, P. A., Meulstee, J. *et al.* Diagnostic and classification criteria for the Guillain-Barre syndrome. *Eur Neurol* **45**, 133-139 (2001).
- 14 Hughes, R. A. and Cornblath, D. R. Guillain-Barre syndrome. *Lancet* **366**, 1653-1666 (2005).
- 15 Lehmann, H. C., Hughes, R. A., Kieseier, B. C. *et al.* Recent developments and future directions in Guillain-Barre syndrome. *J Peripher Nerv Syst* **17 Suppl 3**, 57-70 (2012).
- 16 van Doorn, P. A. Diagnosis, treatment and prognosis of Guillain-Barre syndrome (GBS). *Presse medicale* (2013).
- 17 Walgaard, C., Lingsma, H. F., Ruts, L. *et al.* Early recognition of poor prognosis in Guillain-Barre syndrome. *Neurology* **76**, 968-975 (2011).
- 18 van den Berg, B., Walgaard, C., Drenthen, J. *et al.* Guillain-Barre syndrome: pathogenesis, diagnosis, treatment and prognosis. *Nat Rev Neurol* **10**, 469-482 (2014).
- 19 Van den Bergh, P. Y. and Rajabally, Y. A. Chronic inflammatory demyelinating polyradiculoneuropathy. *Presse medicale* (2013).
- 20 Joint Task Force of the EFNS and the PNS. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of paraproteinemic demyelinating neuropathies. Report of a Joint Task Force of the European Federation of Neurological Societies and the Peripheral Nerve Society—first revision. *J Peripher Nerv Syst* **15**, 185-195 (2010).
- 21 Rajabally, Y. A., Simpson, B. S., Beri, S. *et al.* Epidemiologic variability of chronic inflammatory demyelinating polyneuropathy with different diagnostic criteria: study of a UK population. *Muscle & nerve* **39**, 432-438 (2009).
- 22 Gorson, K. C. An update on the management of chronic inflammatory demyelinating polyneuropathy. *Therapeutic advances in neurological disorders* **5**, 359-373 (2012).
- 23 Querol, L., Rojas-Garcia, R., Casanovas, C. *et al.* Long-term outcome in CIDP patients treated with IVIg: A retrospective study. *Muscle & nerve* (2013).
- 24 Ramchandren, S. and Lewis, R. A. An update on monoclonal gammopathy and neuropathy. *Current neurology and neuroscience reports* **12**, 102-110 (2012).

- 25 Nobile-Orazio, E. Update on neuropathies associated with monoclonal gammopathy of undetermined significance (2008-2010). *J Peripher Nerv Syst* **15**, 302-306 (2010).
- 26 Lunn, M. P. and Nobile-Orazio, E. Immunotherapy for IgM anti-myelin-associated glycoprotein paraprotein-associated peripheral neuropathies. *The Cochrane database of systematic reviews* **5**, CD002827 (2012).
- 27 Leger, J. M., Viala, K., Nicolas, G. *et al.* Placebo-controlled trial of rituximab in IgM anti-myelin-associated glycoprotein neuropathy. *Neurology* **80**, 2217-2225 (2013).
- 28 Niermeijer, J. M., Fischer, K., Eurelings, M. *et al.* Prognosis of polyneuropathy due to IgM monoclonal gammopathy: a prospective cohort study. *Neurology* **74**, 406-412 (2010).
- 29 Muley, S. A. and Parry, G. J. Multifocal motor neuropathy. *Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia* **19**, 1201-1209 (2012).
- 30 Leger, J. M. and Behin, A. Multifocal motor neuropathy. *Current opinion in neurology* **18**, 567-573 (2005).
- 31 Umaphathi, T., Hughes, R. A., Nobile-Orazio, E. *et al.* Immunosuppressant and immunomodulatory treatments for multifocal motor neuropathy. *The Cochrane database of systematic reviews* **4**, CD003217 (2012).
- 32 Van Asseldonk, J. T., Franssen, H., Van den Berg-Vos, R. M. *et al.* Multifocal motor neuropathy. *Lancet neurology* **4**, 309-319 (2005).
- 33 Vlam, L., van der Pol, W. L., Cats, E. A. *et al.* Multifocal motor neuropathy: diagnosis, pathogenesis and treatment strategies. *Nat Rev Neurol* **8**, 48-58 (2012).
- 34 Paolazzi, G., Peccatori, S., Cavatorta, F. P. *et al.* A case of spontaneously recovering multifocal motor neuropathy with conduction blocks (MMNCB) during anti-TNF alpha therapy for ankylosing spondylitis. *Clinical rheumatology* **28**, 993-995 (2009).
- 35 Bulens, C., Vermey, F. H., and van Doorn, P. A. Multifocal motor neuropathy with abrupt onset and spontaneous recovery. *Journal of neurology* **254**, 966-967 (2007).
- 36 Vlam, L., van der Pol, W. L., Cats, E. A. *et al.* Multifocal motor neuropathy: diagnosis, pathogenesis and treatment strategies. *Nature reviews. Neurology* **8**, 48-58 (2012).
- 37 Van den Berg, L. H., Kerkhoff, H., Oey, P. L. *et al.* Treatment of multifocal motor neuropathy with high dose intravenous immunoglobulins: a double blind, placebo controlled study. *J Neurol Neurosurg Psychiatry* **59**, 248-252 (1995).
- 38 Federico, P., Zochodne, D. W., Hahn, A. F. *et al.* Multifocal motor neuropathy improved by IVIg: randomized, double-blind, placebo-controlled study. *Neurology* **55**, 1256-1262 (2000).
- 39 Leger, J. M., Chassande, B., Musset, L. *et al.* Intravenous immunoglobulin therapy in multifocal motor neuropathy: a double-blind, placebo-controlled study. *Brain* **124**, 145-153 (2001).
- 40 Nobile-Orazio, E., Gallia, F., Tuccillo, F. *et al.* Chronic inflammatory demyelinating polyradiculoneuropathy and multifocal motor neuropathy: treatment update. *Current opinion in neurology* **23**, 519-523 (2010).
- 41 Ad Hoc Subcommittee of the American Academy of Neurology AIDS Task Force. Research criteria for diagnosis of chronic inflammatory demyelinating polyneuropathy (CIDP). Report from an Ad Hoc Subcommittee of the American Academy of Neurology AIDS Task Force. *Neurology* **41**, 617-618 (1991).
- 42 Asbury, A. K. and Cornblath, D. R. Assessment of current diagnostic criteria for Guillain-Barre syndrome. *Annals of neurology* **27 Suppl**, S21-24 (1990).
- 43 Miescher, G. C. and Steck, A. J. Paraproteinaemic neuropathies. *Baillieres Clin Neurol* **5**, 219-232 (1996).
- 44 van Schaik, I. N., Bouche, P., Illa, I. *et al.* European Federation of Neurological Societies/Peripheral Nerve Society guideline on management of multifocal motor neuropathy. *European journal of neurology : the official journal of the European Federation of Neurological Societies* **13**, 802-808 (2006).
- 45 Lunn, M. P., Leger, J. M., Merkies, I. S. *et al.* 151st ENMC international workshop: Inflammatory Neuropathy Consortium 13th-15th April 2007, Schiphol, The Netherlands. *Neuromuscular disorders : NMD* **18**, 85-89 (2008).
- 46 Lauria, G., Cornblath, D. R., Johansson, O. *et al.* EFNS guidelines on the use of skin biopsy in the diagnosis of peripheral neuropathy. *European journal of neurology* **12**, 747-758 (2005).
- 47 Lauria, G., Bakkers, M., Schmitz, C. *et al.* Intraepidermal nerve fiber density at the distal leg: a worldwide normative reference study. *J Peripher Nerv Syst* **15**, 202-207 (2010).
- 48 Franssen, H. and Wieneke, G. H. Nerve conduction and temperature: necessary warming time. *Muscle & nerve* **17**, 336-344 (1994).
- 49 Franssen, H., Notermans, N. C., and Wieneke, G. H. The influence of temperature on nerve conduction in patients with chronic axonal polyneuropathy. *Clinical Neurophysiology* **110**, 933-940 (1999).

- 50 Kleyweg, R. P., van der Meche, F. G., and Schmitz, P. I. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle & nerve* **14**, 1103-1109 (1991).
- 51 Dyck, P. J., Davies, J. L., Litchy, W. J. *et al.* Longitudinal assessment of diabetic polyneuropathy using a composite score in the Rochester Diabetic Neuropathy Study cohort. *Neurology* **49**, 229-239 (1997).
- 52 Dyck, P. J., Karnes, J. L., O'Brien, P. C. *et al.* The Rochester Diabetic Neuropathy Study: reassessment of tests and criteria for diagnosis and staged severity. *Neurology* **42**, 1164-1170 (1992).
- 53 Dyck, P. J., Kratz, K. M., Lehman, K. A. *et al.* The Rochester Diabetic Neuropathy Study: design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. *Neurology* **41**, 799-807 (1991).
- 54 Desrosiers, J., Hebert, R., Bravo, G. *et al.* Comparison of the Jamar dynamometer and the Martin vigorimeter for grip strength measurements in a healthy elderly population. *Scandinavian Journal of Rehabilitation Medicine* **27**, 137-143 (1995).
- 55 Mathiowetz, V. in *Muscle strength testing. Instrumented and non-instrumented systems* (ed L. R. Amundsen) 163-177 (Churchill Livingstone, 1990).
- 56 Peters, M. J., van Nes, S. I., Vanhoutte, E. K. *et al.* Revised normative values for grip strength with the Jamar dynamometer. *J Peripher Nerv Syst* **16**, 47-50 (2011).
- 57 Merckies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 58 Booth, J. and Young, M. J. Differences in the performance of commercially available 10-g monofilaments. *Diabetes Care* **23**, 984-988 (2000).
- 59 Paisley, A. N., Abbott, C. A., van Schie, C. H. *et al.* A comparison of the Neuropen against standard quantitative sensory-threshold measures for assessing peripheral nerve function. *Diabetic Medicine* **19**, 400-405 (2002).
- 60 Farrar, J. T., Young, J. P., Jr., LaMoreaux, L. *et al.* Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* **94**, 149-158 (2001).
- 61 Maxwell, C. Sensitivity and accuracy of the visual analogue scale: a psycho-physical classroom experiment. *Br J Clin Pharmacol* **6**, 15-24 (1978).
- 62 Graham, R. C. and Hughes, R. A. A modified peripheral neuropathy scale: the Overall Neuropathy Limitations Scale. *J Neurol Neurosurg Psychiatry* **77**, 973-976 (2006).
- 63 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 64 Vickrey, B. G., Hays, R. D., and Beckstrand, M. Development of a health-related quality of life measure for peripheral neuropathy. *Neurorehabil Neural Repair* **14**, 93-104 (2000).
- 65 Darweesh, S. K., Polinder, S., Mulder, M. J. *et al.* Health-related quality of life in Guillain-Barre syndrome patients: a systematic review. *J Peripher Nerv Syst* **19**, 24-35 (2014).
- 66 Bergner, M., Bobbitt, R. A., Carter, W. B. *et al.* The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care* **19**, 787-805 (1981).
- 67 Jacobs, H. M., Luttik, A., Touw-Otten, F. W. *et al.* [The sickness impact profile; results of an evaluation study of the Dutch version]. *Nederlands Tijdschrift Voor Geneeskunde* **134**, 1950-1954 (1990).
- 68 Aaronson, N. K., Muller, M., Cohen, P. D. *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *Journal of Clinical Epidemiology* **51**, 1055-1068 (1998).
- 69 Erdman, R. A., Passchier, J., Kooijman, M. *et al.* The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects. *Psychological Reports* **72**, 1027-1035 (1993).
- 70 Guyatt, G. H., Feeny, D. H., and Patrick, D. L. Measuring health-related quality of life. *Annals of Internal Medicine* **118**, 622-629 (1993).
- 71 Hyland, M. E. A brief guide to the selection of quality of life instrument. *Health Qual Life Outcomes* **1**, 24 (2003).
- 72 Sprangers, M. A., Moinpour, C. M., Moynihan, T. J. *et al.* Assessing meaningful change in quality of life over time: a users' guide for clinicians. *Mayo Clinic Proceedings* **77**, 561-571 (2002).
- 73 Saxena, S., Carlson, D., and Billington, R. The WHO quality of life assessment instrument (WHOQOL-Bref): the importance of its items for cross-cultural research. *Quality of Life Research* **10**, 711-721 (2001).

- 74 Trompenaars, F. J., Masthoff, E. D., Van Heck, G. L. *et al.* Content validity, construct validity, and reliability of the WHOQOL-Bref in a population of Dutch adult psychiatric outpatients. *Quality of Life Research* **14**, 151-160 (2005).
- 75 The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. *Health Policy* **16**, 199-208 (1990).
- 76 Ware Jr, J. E., Snow, K. K., M., K. *et al.* *SF-36 health survey. Manual and interpretation guide.*, (The health institute, New England Medical Center, 1997).
- 77 Ware Jr, J. E., Kosinski, M., and Keller, D. S. *SF-36 physical and mental health summary scales: a user's manual.* (The Health Assessment Lab, New England Medical Center, 1994).
- 78 Law, M., Baptiste, S., McColl, M. *et al.* The Canadian occupational performance measure: an outcome measure for occupational therapy. *Canadian Journal of Occupational Therapy. Revue Canadienne d'Ergotherapie* **57**, 82-87 (1990).
- 79 Law, M., Polatajko, H., Pollock, N. *et al.* Pilot testing of the Canadian Occupational Performance Measure: clinical and measurement issues. *Canadian Journal of Occupational Therapy. Revue Canadienne d'Ergotherapie* **61**, 191-197 (1994).
- 80 Van Nes, S. I. *Improving and standardising assessment of patients with immune-mediated neuropathies: Peripheral Neuropathy outcome measures Standardisation study (PeriNomS study – part 1)* PhD thesis, Erasmus MC: University Medical Center Rotterdam, (2012).
- 81 Hughes, R., Bensa, S., Willison, H. *et al.* Randomized controlled trial of intravenous immunoglobulin versus oral prednisolone in chronic inflammatory demyelinating polyradiculoneuropathy. *Annals of neurology* **50**, 195-201 (2001).
- 82 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Connecting impairment, disability, and handicap in immune mediated polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **74**, 99-104 (2003).
- 83 Merkies, I. S., van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *Journal of neurology, neurosurgery, and psychiatry* **81**, 1194-1199 (2010).
- 84 Vanhoutte, E. K., Faber, C. G., and Merkies, I. S. 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscular disorders : NMD* (2013).

Chapter 2

# **Rasch-ionale for neurologists**



## Abstract

Outcome measures are considered the most important tools to monitor patients' outcome in both clinical and research settings. Measuring the clinical state of patients is a fundamental part of our daily clinical practice and research that sometimes has been taken for granted. Whereas most physicians accept that it may take over 12 years to become a medical specialist, virtually no medical expert considers the efforts needed to understand the science of proper assessments for clinical follow-up studies and trial design and conduct. In peripheral neuropathies, there are several measurement scales available, but most of these are at the ordinal level. This paper will systematically address the types of data available (nominal, ordinal, interval, and ratio data) in terms of their strengths and weaknesses. The differences between classical test theory-based and modern test methods-based outcome measures will be addressed with emphasis on Rasch methodology. Various steps will be highlighted as part of the evaluation and construction of outcome measures using the Rasch method, with the aim to increase the knowledge and utility of this technique. Rasch-built outcome measures should be used for future studies in inflammatory neuropathies and their method of construction could be easily extrapolated to other neurological illnesses.

## Introduction

Assessing outcome in clinical studies would have been much easier if there was a metric for every quality of interest (e.g. mobility, fatigue, depression, pain, daily/social functioning, etc) with a fixed unit throughout the metric's range (such as measuring height or weight). In that case, trial design and clinical follow-up studies would have been much easier and this would have facilitated the routine use of standardized outcome measures. Unfortunately, most qualities of interest in the medical field are less tangible and do not have a fixed internationally standardized ruler for their assessments. Surrogate outcome measures are often being applied and are often considered the best second option to collect information about a patient in both clinical practice and research setting. Measurement is such a fundamental part of our daily clinical activities that we sometimes may take the construction and use of outcome measures for granted. While everyone accepts that it generally takes over 12 years to become a medical specialist in some fields like neurology, virtually no medical expert considers the efforts to understand the science of proper assessments and use in clinical trial design and conduct. A review of 713 studies on [www.clinicaltrials.gov](http://www.clinicaltrials.gov) (March 10<sup>th</sup> 2014; using the terms "peripheral neuropathy" and "randomized") showed that 85% used outcome measures at the ordinal level of assessment, despite the shortcomings of such scales that have been addressed in the literature for decades.<sup>1-4</sup> Types of collected data also should be part of the educational arsenal for clinicians, increasing their awareness regarding their (dis)advantages regarding when and how they are used.<sup>3</sup> Also, new methodological techniques for the development of outcome measures, like the Rasch model and the Item Response Theory (IRT) have been increasingly recognized, although not generally applied.<sup>5-7</sup> A plausible explanation would be that these methods seem to have a complex mathematical background, and therefore less accessible to most physicians, including neurologists. Nevertheless, the superiority of modern clinimetric approaches compared to classical test methods is now established and therefore should be applied in future studies for designing proper outcome measures.<sup>1-4,8</sup>

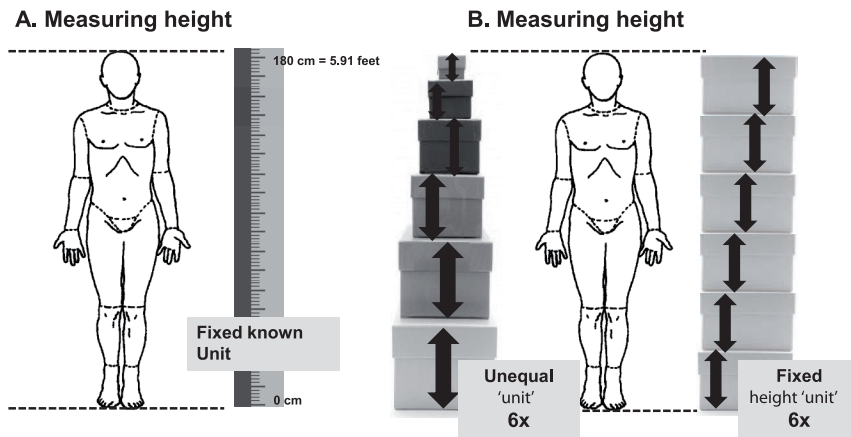
The current paper has the following purposes: a) to systematically address available types of data; b) to demonstrate the differences between classical versus modern test theories, emphasizing the advantages of modern clinimetric methods; and, c) to highlight the various steps needed in evaluating and constructing outcome measures using the Rasch method, which is applied as an example in the inflammatory neuropathy field and may be extrapolated to other neurological areas as well.

## Measurement and types of data

Measuring the height of a patient is generally performed without any difficulty, using a ruler with a fixed unit (in Europe: centimeters, in Anglo-Saxon community: foot/inch;

see Figure 1A). However, in health care, it is often necessary to measure qualities such as pain, fatigue or disability, where there is no ruler with a fixed unit available. Such qualities are nevertheless generally assessed using constructed surrogate outcome measures that are often composed by a set of items or questions (Figure 1B).

**Figure 1.** Measuring using a ruler versus measuring using a surrogate measure

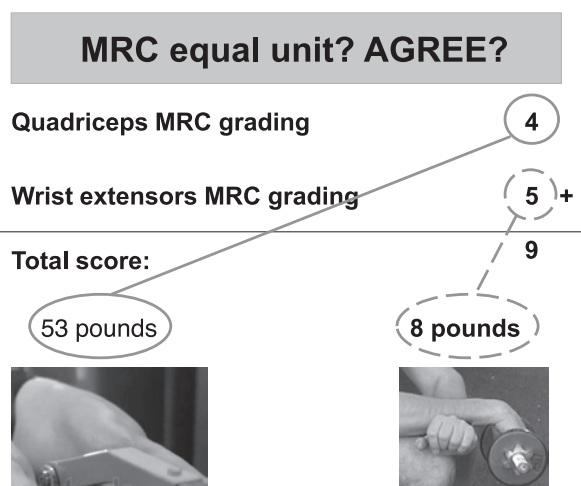


**Legend to Figure 1:** A: measuring a patient's height using a linear measure with known fixed units; B: measuring a patient's quality if there is not a ruler with a fixed unit for this. The boxes symbolize the various items of a composite surrogate measure.

In science, "measure" involves numbers that can be used in calculations. Even after multiplying and dividing, the numbers should maintain their values ("known unit"). Most clinical trials performed thus far in inflammatory neuropathies have collected raw data from a set of items or questions that have not been measured in this way.<sup>9,10</sup> Whether the collected numbers can be used in calculations depends on the type of data in the scale. For example, data at the nominal level consist of numbers or labels with unknown magnitude. It is therefore not possible to determine which category is greater or less than another. Examples of these are gender, blood type, religion, etc. Data measured at the ordinal level can be ordered or ranked, but the difference between the categories is unknown or unequal (e.g., level of disability). Thus by ranking data the only information that can be obtained is whether a value is greater or less relative to the other. Since the distance between the numbers assigned to the different categories is highly unlikely to be equal, sum scores cannot be used. Also, it is not possible to compare different multi-item composite scales measuring the same trait.<sup>1-4,10</sup> For example, the ODSS<sup>11</sup> and the Rankin disability scale<sup>12</sup> are both disability scales, but their outcomes cannot be compared. In addition, using the Medical Research Council (MRC) grading system to examine strength at bedside assumes a fixed unit for the assessment of strength in the various muscle groups.<sup>13,14</sup> However, the

MRC grading system is a descriptive categorical measure, thus the data collected are at the ordinal level with no intrinsic numerical value. Even if there was a fixed unit, one cannot understand how a patient with an MRC grade 4 for the quadriceps muscle, is able to push 53 pounds during gym exercise, while the same patient has MRC grade 5 (normal strength) for the wrist extensors only lifts 8 pounds with this group of muscles (Figure 2). The inequality of the MRC grading system has already been extensively demonstrated by comparing the distribution of the grades with a dynamometer (data at the ratio level) for assessing strength.<sup>15-19</sup>

**Figure 2.** Medical Research Council (MRC) grading system in relation to the amount of pounds that can be lifted.



**Legend to Figure 2:** This is an example of the MRC grades seen in a patient with chronic inflammatory demyelinating polyradiculoneuropathy in two muscle groups. The MRC grade in the wrist extensors is 5, this muscle group can only lift 8 pounds of weight, while the MRC grade being 4 in the quadriceps is still able to push/lift 53 pounds, indicating unit inequality of the MRC grading system

When the distance between two successive categories is known, measurement is at the interval level and meaningful calculations can be made. Finally, when a non-arbitrary zero value indicating the lack of whatever you try to measure is present, a ratio scale is created (e.g. length, weight). Conventional (or parametric) statistics (e.g. t-test) can only be applied at the interval or ratio level, where a known unit of difference is present between two successive categories.<sup>20</sup>

In addition, physicians and researchers should be aware that there are four domains of measurement: the level of pathology, impairment, activity and participation, and the concept of quality of life.<sup>21,22</sup> Outcome measures should solely represent one of these levels using only items at that particular domain, since the findings at one domain,

for example impairment measures (like weakness) may be correlated with activity limitations in daily and social functioning (weakness leading to disability).<sup>21-23</sup> Having items representing various levels of assessing outcome may inflate the final sum score and provide improper results.<sup>6</sup>

## Models to create outcome measures

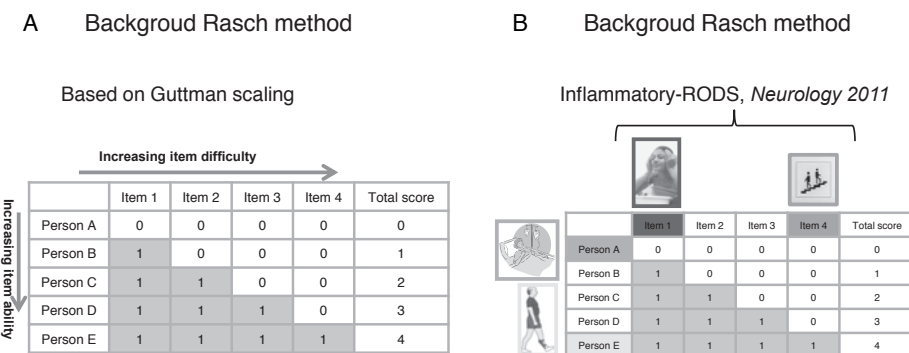
A mathematical model is required to assess a particular patient's quality of interest. Such a model is a mathematical representation of reality and can be used to create a surrogate outcome measure in the absence of a known ruler with a fixed unit.<sup>24</sup> The Classical Test Theory (CTT) is the most widely used model, based on the assumption that the true score is the sum of the obtained score (raw data) and random error.<sup>4,24</sup> However, the magnitude of random error is unknown. Outcome measures based on the CTT are always ordinal in nature and have limitations as discussed previously.<sup>1,2,4</sup> In brief, items as part of a CTT-ordinal based metric are arbitrarily taken generally using a Likert-type form of response options; in addition, it is assumed that all parts of such a metric have equal contribution to the scale, which is highly unlikely. The various weights of each item are generally unknown. Often, sum scores have been constructed assuming linearity of such a measure, which is also highly unlikely.<sup>1,2,4,24,25</sup> Thus a difference in the scores will not have the same meaning throughout the range of the scale of interest. These observations may lead to wrong conclusions in intervention trials, such as false positive or negative results.

There are two modern techniques to transform outcome measures to the interval level: the Rasch model and the Item Response Theory (IRT), which have the same mathematical background. The IRT approaches include additional model parameters to reflect the patterns observed in the data.<sup>9,20,26</sup> On the other hand, according to the Rasch approach, the data should fit the Rasch model, before any reliable claim about the presence of a trait can be made. Therefore misfitting responses require further examination to explore the reason for the misfit, and may be excluded from the data set if one can explain substantively why they do not address the latent trait.<sup>5,10,24,27</sup> When the expectations of the model are fulfilled, interval measurement can be obtained. Rasch analysis should be used when there is a need to use a set of items in order to create sum scores. In scale development, Rasch can be used to create a new unidimensional outcome measure free from item bias. Also, the scientific properties of existing scales can be examined using Rasch and item banks can be constructed.<sup>6,28</sup> Another major advantage of the Rasch model is that it overcomes the problem of missing data by providing calculations of these as part of the modeling.<sup>26,29</sup>

# Background of the Rasch model

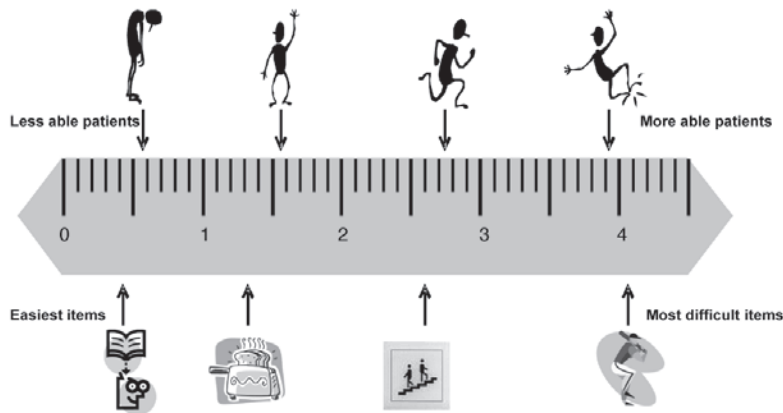
In 1960 Georg Rasch stated that: “a person having a greater ability than another person should have the greater probability of solving any item or type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one”.<sup>5</sup> In other words, patients with a higher ability (less ill) should have a greater chance of obtaining a higher score on any item/task of interest. Thus, the probability of confirming an item depends on the difference between the ability (“quality of interest”) of the person and the difficulty of the item.<sup>5,6,26</sup> For example, a patient with severe disability due to chronic inflammatory demyelinating polyradiculoneuropathy (CIDP) is expected to have more difficulty climbing a flight of stairs than a less disabled patient.<sup>30</sup> Based on the so-called Guttman scaling, people can be arranged in such a manner that if a person is able to perform a particular task, this individual should also have a high probability to complete tasks with a lower rank order. Accordingly, items can be ranked from most easy to most difficult (Figure 3).<sup>31</sup> Thereafter both the “person ability” and the “item difficulty” can be placed on the same mutual ruler (Figure 4).

**Figure 3.** Example of a Guttman scaling



**Legend to Figure 3:** The Guttman scale orders the difficulty of items and ability of persons in a matrix (A). An example of a Rasch-built scale is provided in Figure 3B; the inflammatory-RODS showing the item “able to wash you face” being one of the easiest items compared to the “ability to walk a flight of stairs”.<sup>30</sup>

After ordering the obtained ordinal data into a Guttman table, the model subsequently estimates the person ability and item difficulty separately and independently from each other and expresses these as a log odds ratio, or Logits (Appendix 1; the dichotomous Rasch formula explained).<sup>20,24</sup> After fulfilling additional requirements of the Rasch model, interval measurement on a logits scale can be obtained from ordinal measures.<sup>6</sup>

**Figure 4.** Interval ruler

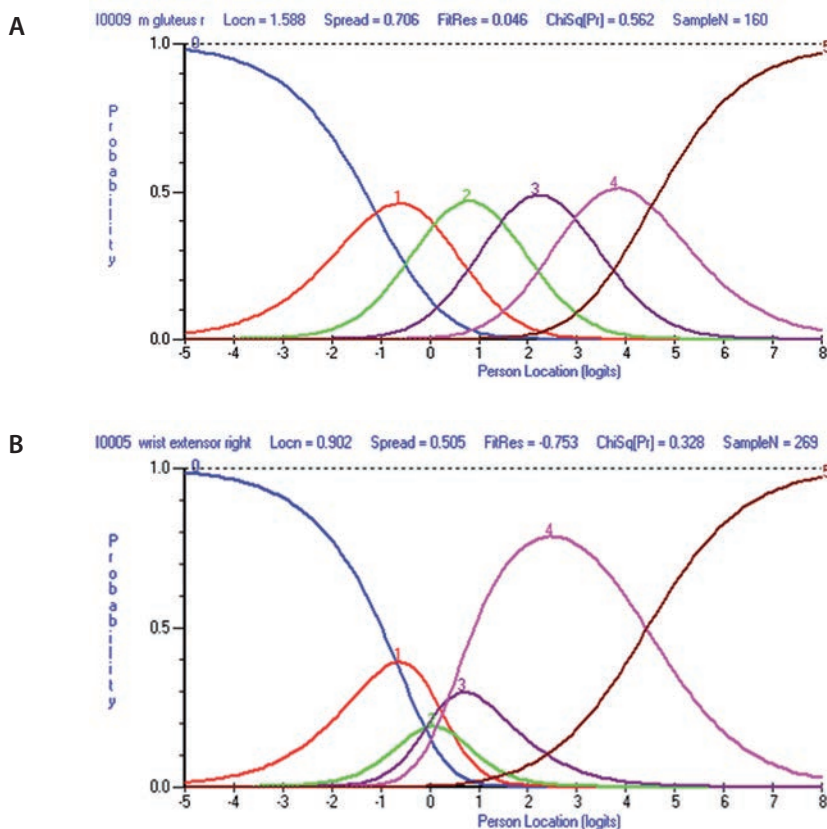
**Legend to Figure 4:** Using interval instead of ordinal measures provides a true reflection of disease impact, differences between individuals and groups, and treatment effects. The corresponding weights ("location") of the items and patients also are calculated by the model depending on the interaction between the items and the patients.

## Rasch model requirements

To obtain a scale at the interval level, all items and persons as part of the model need to fulfill several Rasch methodology requirements, or checkpoints. Most requirements of the Rasch model are based on the assumption that the obtained raw data should not deviate significantly from the ideal situation, that is, what was estimated by the Rasch model.<sup>6,20,24</sup> The Rasch unidimensional measurement model (RUMM) is a computer program that can be used to check whether the data fit the model expectations, and items or patients not fulfilling these requirements should be removed or subjected to re-adjustments to fit the model's requirements.<sup>29,32</sup> The following steps are checkpoints that should be accomplished to achieve model fit.

### Parameterization of the Rasch model

The Rasch model contains several specifications, and which one should be used depends on the number of response options. When a set of items contain only two response options, the dichotomous model is used.<sup>6</sup> However, when more than two response options are present a 'threshold' parameter is applied (Figure 5).<sup>33</sup> A threshold is the point between two adjacent categories where either response is equally probable. The Partial Credit Model is an extension of the threshold parameter model and makes no assumptions about the equality of threshold locations relative to each item.<sup>33</sup> The Partial Credit Model is used as default in RUMM. However, when the Likelihood Ratio test probability is not significant, the Rating Scale model could be used.<sup>32</sup>

**Figure 5.** Examples of ordered and disordered thresholds

**Legend to Figure 5:** If patients cannot discriminate properly between the response options, “reversed or disordered thresholds” will occur. An example of ordered thresholds is given in figure A, whereas figure B shows disordered thresholds.

### Targeting and person-item distribution

Rasch always places the mean of the items around zero logits and this represents the average difficulty of a scale. The mean of persons can be compared to the item mean of zero and this provides an indication of the targeting of the scale. When the persons mean deviates from zero, the targeting might be at risk. If this value is positive the items may have been too easy for the person, and a negative value suggests the opposite.<sup>6</sup> Furthermore, the spread of the persons should be wide enough. This can be examined using the item-person distribution. In the ideal situation, the responses should preferably have no floor or ceiling effects. Items and persons should be distributed in such a way that the items can be distinguished among persons. Persons in the middle of the test have the smallest error of measurement.<sup>29</sup>



## Reliability

Internal reliability can be estimated, based on the Person Separation Index (PSI). The PSI provides an indication about the ability to discriminate between different groups. A value exceeding 0.7 indicates the ability to identify at least two groups of patients.<sup>28</sup>

## Sample size

Preferably, a minimum of 150 patients, and preferably 250, would be required to have 99% confidence that the item calibrations remain stable within 0.5 logits.<sup>34</sup>

## Unidimensionality

Although human behavior is complex in nature, measurement requires that only one aspect at the time is being measured. The Rasch model requires that all items summed together form a unidimensional outcome measure. For example, when assessing levels of disability, there should be no questions about other aspects, such as fatigue or quality of life. Exploring a wider range of experiences does not only add information about the trait of interest (e.g., disability), but addresses other aspects of behavior and thereby would create multidimensional scales. The outcome of these scales is not interpretable since it does not only provide (an unknown amount of) information about the latent trait, but also an unknown amount of information about other aspects. When the response to an item depends on a factor other than the person's ability, unidimensionality is violated.<sup>28</sup>

## Fit analysis

The method requires that all raw data do not deviate significantly from the model's expectation. For example, patients who are able to perform "difficult" but not "easy" items are reason for concern. Item fit also needs to be examined. The Rasch model estimates of person ability and item difficulty represent the ideal situation where the data are forced to be linear and unidimensional. There are several statistics available to determine whether the raw data deviate from the expected 'ideal' situation. Fit statistics, considered as forms of checkpoints, give an indication whether the quality is sufficient to allow the output to be interpreted as an interval measurement.<sup>20,28</sup> Also, when targeting is not optimal, fit statistics should be interpreted with some caution. There are two general categories for detecting misfit: overall misfit, using the entire response matrix, and the individual fit (examining all items and all persons individually).<sup>28,33</sup>

## Overall fit statistics

In RUMM a summary of statistics is provided. Three overall fit statistics are generally considered. Item fit statistic places the mean item difficulty arbitrarily at 0 logits. The person mean is estimated in relation to the item difficulty. Residuals are calculated using the difference between what is observed (raw data, responses) and what is expected by the model (the probabilities). Both for the persons and the items, the

overall mean residual values can be calculated. Residuals are transformed into a z-score, indicating the number of standard deviations a given observation is above or below the mean.<sup>10,28,33</sup> When data fit the model, a mean of 0 and a standard deviation of 1 can be expected. As a rule of thumb, a fit residual standard deviation greater than 1.4 indicates misfit.<sup>6,28</sup> The next overall statistic is the item-trait interaction statistic that reflects the fit of the observed data to the model's estimates and is represented by the chi-square. It is calculated from a comparison of the observed overall performance of each trait-group (class interval) on the item with its expected performance. Therefore, it gives an indication of the degree of invariance of each of the items and how the items function across the trait. A significant chi-square means overall item misfit, indicating multidimensionality.<sup>28</sup>

### Individual person and item fit

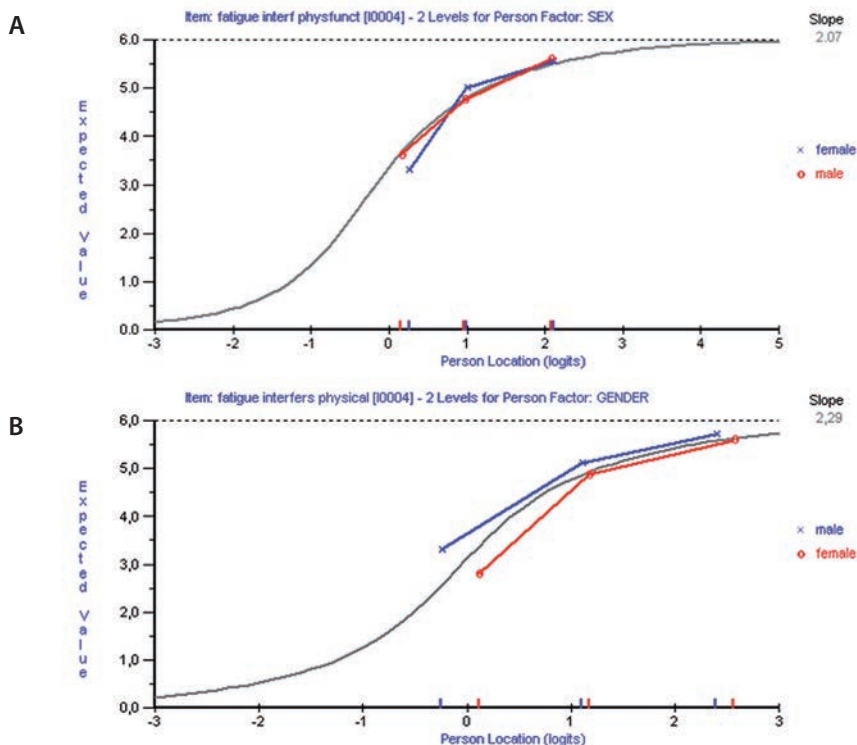
Besides the overall fit residuals, the individual residuals can be calculated. Although the residual cannot achieve zero, individual residual values outside  $\pm 2.5$  should be examined. A significant item chi-square indicates significant deviation of the model. Item fit residuals with high positive values are underdiscriminating and suggest poor fit to the model. High negative residuals are overdiscriminating and do not add extra information. The individual item chi-square should not be significant; as a significant chi-square indicates that the item deviates from the model. Items or persons showing misfit to the model should be examined. As a final solution items or persons can be deleted to improve the scale construct.<sup>6,28</sup>

### Working of response categories

For polytomous data, the point between two adjacent response categories where the probability on either response is equal is referred to as a threshold (Figure 5). Disordered thresholds are a source of misfit and occur when respondents use the response options inconsistently (Figure 5). Persons can experience difficulties discriminating consistently between response options, when the labeling of the options is confusing, open to misinterpretation or when there are too many options. In general, it has been argued that adults are not able to discriminate between more than 4 response options.<sup>35</sup> Disordered thresholds usually can be resolved by collapsing categories. The number of patients in each class interval should always be checked before collapsing categories.<sup>28</sup>

### Differential Item Functioning (DIF)

DIF occurs when persons in different subgroups (e.g. based on age, gender, nationality or underlying disorder) with equal ability levels respond differently to an individual item. In the item characteristic curve (ICC), the proportion of persons with the same ability who confirm a particular item should remain the same irrespective of any subgroup (Figure 6). Items that do not yield the same item response function for 2 or more groups display DIF (Figure 6) and are violating the requirement of unidimensionality.<sup>6,28,33,36</sup>

**Figure 6.** Differential Item Functioning (DIF)

**Legend to Figure 6:** A scale should always work in the same way irrespective of which personal factor (e.g., gender, country) is being examined. If this is the case, there is no DIF present (6A). If patients with equal ability levels respond systematically differently to an item, this item shows DIF, and will violate the requirement for unidimensionality (6B). DIF: differential item functioning. ICC: Item Characteristic Curve.

There are two types of DIF, uniform and non-uniform DIF. Uniform DIF occurs when the persons in a particular subgroup respond consistently and systematically differently, across the whole range of the ability. Non-uniform DIF occurs when the difference varies across the ability levels. The statistical test used for detection of DIF is an ANOVA of the person-item deviation residuals with person factors (e.g. age, gender, nationality, disease duration) and class intervals (e.g. group along the trait) as factors. A summary of DIF is presented in the ANOVA table of the person-item deviation residuals for each person factor and class interval using the RUMM program. Uniform DIF is visualized by a significant p-value on DIF (item) effect, and non-uniform DIF by a significant Item-by class interval p-value. With many tests of fit, some will be significant just by chance; therefore the Bonferroni correction should be applied.<sup>37</sup> When DIF is present in one group, artificial DIF can appear in another group.<sup>38,39</sup> If there is more than one item showing DIF, the first thing to do is to split or remove the item with the highest mean

square, since this will be the item displaying real DIF. Items displaying artificial DIF should then disappear. Uniform DIF can be resolved by splitting the item on person factor, or by removing the item. For non-uniform DIF removal of the item is often the only solution.<sup>6,28</sup>

### Response dependency

The reliability can be inflated and the final scale's score can be forced in a particular direction when the responses to items in an outcome measure are related to each other. In other words, affirming one item should not automatically lead to the affirmation of another item (e.g., a person stating he is able to walk one kilometer, will also be able to walk 100 meters). Local dependency can be discovered through the residual correlation matrix. Combining items in a subtest or eliminating items solves the problem of local dependency and improves model fit. Residual correlations above 0.3 needs to be resolved by combining the items in a subtest or by deleting items.<sup>6</sup>

### Responsiveness

Sensitivity to change, or responsiveness, is defined as the ability of an instrument to detect important changes over time and is a required quality for any instrument to be used in clinical trials and research designs. The sensitivity to change can be tested according to both group-level and individual-level approaches. The group-level approach consists of computing different responsiveness indices in groups of patients (paired t-test, effect size and standardized response mean). The individual approach consists of taking into account the standard error of measurement associated with the patient's score obtained during both evaluation. It is important to underline here that any Rasch-built scale provides the individual's measures with associated standard error of measurement. A statistic for each patient can be computed to test the extent to which their personal score has changed. The individual-level approach presents a considerable advantage compared to the group-level approach since it can report whether the score level of a particular patient has significantly improved or decreased. Indeed, meaningful change for groups of patients may not have the same significance for individuals. Consequently the individual-level approach provides clinicians an alternative method of drawing conclusions from group results to individuals. Results can henceforth be interpreted patient by patient.

### Conclusion

The Rasch model can be seen as the ideal response pattern, where persons with high abilities should have a higher probability of receiving a better score on any item compared to persons with a lower ability. Also, any person should always have a greater probability of receiving a higher score on an easier item than on a more difficult one.

In the model, person ability and item difficulty are estimated separately. The difference between the ability and the difficulty should not deviate significantly from the model and various checkpoints are being used to monitor this. All data should fulfill the Rasch model's expectations like good item and person statistical fit, threshold ordering, no item bias or local dependency and demonstrating unidimensionality. Only after these criteria are satisfied will the measurement be at the interval level on a logits scale and suitable for conventional statistics. Only then, meaningful sum scores and changes in scores can be measured properly.

## References

- 1 Grimby, G., Tennant, A., and Tesio, L. The use of raw scores from ordinal scales: time to end malpractice? *Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine* **44**, 97-98 (2012).
- 2 Merbitz, C., Morris, J., and Grip, J. C. Ordinal scales and foundations of misinference. *Archives of physical medicine and rehabilitation* **70**, 308-312 (1989).
- 3 Stevens, S. S. On the theory of scales of measurement. *Science* **103**, 677-680 (1946).
- 4 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-59 (2006).
- 5 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (1960).
- 6 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 7 Hays, R. D., Morales, L. S., and Reise, S. P. Item response theory and health outcomes measurement in the 21st century. *Medical care* **38**, 1128-42 (2000).
- 8 Svensson, E. Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine* **33**, 47-48 (2001).
- 9 Hobart, J. C., Cano, S. J., Zajicek, J. P. *et al.* Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet neurology* **6**, 1094-1105 (2007).
- 10 Tesio, L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine* **35**, 105-115 (2003).
- 11 Merckies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Clinimetric evaluation of a new overall disability scale in immune mediated polyneuropathies. *J Neurol Neurosurg Psychiatry* **72**, 596-601 (2002).
- 12 van Swieten, J. C., Koudstaal, P. J., Visser, M. C. *et al.* Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* **19**, 604-607 (1988).
- 13 Vanhoutte, E. K., Faber, C. G., van Nes, S. I. *et al.* Modifying the Medical Research Council grading system through Rasch analyses. *Brain: a journal of neurology* **135**, 1639-1649 (2012).
- 14 Medical Research Council. *Medical Research Council. Aids to the investigation of the peripheral nervous system.*, (Her Majesty's Stationary Office, 1943).
- 15 Brandsma, J. W., Schreuders, T. A., Birke, J. A. *et al.* Manual muscle strength testing: intraobserver and interobserver reliabilities for the intrinsic muscles of the hand. *Journal of hand therapy: official journal of the American Society of Hand Therapists* **8**, 185-190 (1995).
- 16 Cuthbert, S. C. and Goodheart, G. J., Jr. On the reliability and validity of manual muscle testing: a literature review. *Chiropractic & osteopathy* **15**, 4 (2007).
- 17 MacAvoy, M. C. and Green, D. P. Critical reappraisal of Medical Research Council muscle testing for elbow flexion. *The Journal of hand surgery* **32**, 149-153 (2007).
- 18 Merlini, L. Measuring muscle strength in clinical trials. *Lancet neurology* **9**, 1146; author reply 1146-1147 (2010).
- 19 van der Ploeg, R. J., Oosterhuis, H. J., and Reuvekamp, J. Measuring muscle strength. *Journal of neurology* **231**, 200-203 (1984).
- 20 Bond, T. G. and Fox, C. M. *Applying the Rasch model: fundamental measurement in the human sciences*. 2nd edn, (Lawrence Erlbaum Associates Publishers, 2007).
- 21 World Health Organization. *ICF: international classification of functioning, disability and health*. (World Health Organization, 2001).
- 22 Aaronson, N. K. Quality of life: what is it? How should it be measured? *Oncology (Williston Park)* **2**, 69-76, 64 (1988).
- 23 Merckies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Connecting impairment, disability, and handicap in immune mediated polyneuropathies. *J Neurol Neurosurg Psychiatry* **74**, 99-104 (2003).
- 24 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* **13**, iii, ix-x, 1-177 (2009).

- 25 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of clinical epidemiology* **49**, 711-717 (1996).
- 26 Wright, B. D. and Mok, M. M. C. in *Introduction to Rasch measurement* Ch. 1, 1-24 (Jam Press, 2004).
- 27 Massof, R. W. The measurement of vision disability. *Optometry and vision science : official publication of the American Academy of Optometry* **79**, 516-552 (2002).
- 28 Pallant, J. F. and Tennant, A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *The British journal of clinical psychology / the British Psychological Society* **46**, 1-18 (2007).
- 29 Andrich D, L. G., Sheridan BE. (RUMM Laboratory, 2004).
- 30 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 31 Altman, D. G. *Practical statistics for medical research*. 1st edn, (Chapman and Hall, 1991).
- 32 Andrich, D., Lyne, A., Sheridan, B. *et al.* *Rasch Unidimensional Measurement Models (RUMM2020 Version 4.0)*. (2003).
- 33 Tennant, A., Penta, M., Tesio, L. *et al.* Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care* **42**, 137-48 (2004).
- 34 Linacre, J. M., Heinemann, A. W., Wright, B. D. *et al.* The structure and stability of the Functional Independence Measure. *Archives of physical medicine and rehabilitation* **75**, 127-132 (1994).
- 35 Penta, M., Tesio, L., Arnould, C. *et al.* The ABILHAND questionnaire as a measure of manual ability in chronic stroke patients: Rasch-based validation and relationship to upper limb impairment. *Stroke* **32**, 1627-1634 (2001).
- 36 Holland, P. W., Wainer, H., and Educational Testing Service. *Differential item functioning*. (Lawrence Erlbaum, 1993).
- 37 Bland, J. M. and Altman, D. G. Multiple significance tests: the Bonferroni method. *BMJ* **310**, 170 (1995).
- 38 Andrich, D. and Hagquist, C. Real and artificial differential item functioning. *Journal of educational and behavioral statistics* (2011).
- 39 Brodersen, J., Meads, D., Kreiner, S. *et al.* Methodological aspects of differential item functioning in the Rasch model. *Journal of medical economics* **10**, 309-324 (2007).
- 40 Smith, E. V. J. and Smith, R. M. *Introduction to Rasch Measurement: Theory, Models, and Applications*. 1-23 (JAM Press, 2004).

## Appendix 1. The (dichotomous) Rasch formula explained.

To explain the Rasch model, an imaginary test with 9 items (small letters: a through i) and two response options (0 = unable, 1 = able) will be used. The scores of 9 persons (capital letters A through I) are provided and ordered according the sum scores of both the persons and the questions (Table 1).

Thus far an ordered (Guttman) matrix of the raw scores has been created.<sup>31</sup> Based on the sum scores, the percentage of success can be calculated for both the items and the persons by dividing the sum score of each completed item(s)/person by the total number of items,  $n/N\%$  (example: patient A was able to perform 8 (sum score: 8) of the 9 items, thus having 89% of success).

These percentages are still at the ordinal level. Furthermore, probabilities of 0% or 100% are unrealistic, since there is always a small chance that a person is unable to perform a task. Therefore, the probability of success should be expressed as its ratio: odds favoring a positive score are expressed as the ratio of the probability that an event will happen to the probability that it will not happen.

Thus,  $\text{odds} = p/(1-p)$ , where  $p$  = percentage of success. In case of patient A:  $p=0.89$  leading to an odd:  $0.89/(1-0.89) = 8.09$ . The odds can be seen as the likelihood that success will happen (Table e-1, row and column odds).

At this point it is known which person has more correct answers than another person, and it is known what item is more difficult than another one. However, it is unknown how much more able or difficult the person or the item is. "How much" is not a ratio question; but a matter of difference.<sup>26</sup> By taking the natural logarithm of the odds, an estimate of the ability of persons ( $B_n$ ) and an estimate of the difficulty of items ( $D_i$ ) is provided on interval scale, namely a log odds ratio or logits scale (Table e-1, row and column logits).

The Rasch model states that the probability of any outcome occurring only depends on the difficulty of an item and the ability of a person. Thus, the probability for any person successfully performing any item/task is governed only by the difference between the ability of a person ( $B_n$ ) and the item difficulty ( $D_i$ ), i.e. ( $B_n - D_i$ ). Thus, the probability of receiving the correct answer  $P(X_1)$  can be expressed as the ability of a person ( $B_n$ ) minus the difficulty of an item ( $D_i$ ), thus:  $P(X_1) = B_n - D_i$ . When the ability of a person exceeds the difficulty of an item ( $B_n - D_i > 0$ ) the probability of success is more than 50%. When the ability of a person is lower than the difficulty of the item ( $B_n - D_i < 0$ ) the chance of success is less than 50 percent. When the ability of a person is equal to the difficulty of an item ( $B_n - D_i = 0$ ) the probability of success is 50%.

In theory,  $B_n - D_i$  can vary from minus infinity to plus infinity. However, since the Rasch model only addresses the probability of achieving a particular score, the differences can only range from 0 to 1 (0% to 100%). Expressing the odds ratio of the probability of success as a natural logarithm will create a scale approaching linearity and will range between 0 and 1. In Table e-1 row and column logits are provided.



Thus, the chance on getting the correct answer,  $P(X=1)$  can be obtained by taking the natural logarithm of the odds ratio of patient's ability minus the difficulty of the item:  $P(X=1) = \text{Ln}(\text{Probability of correct answer}1 - \text{probability of correct answer})$ .

Therefore,  $(Bn-Di) = \text{Ln } Pni1 - Pni$ ; where  $Pni$  indicates the probability on success,  $Bn$  the ability of person  $n$  and  $Di$  the difficulty of item  $i$ . Since we are interested in the value of  $Pni$ , we should get rid of the  $\text{Ln}$ , and therefore rewrite the formula to:

$Pni = \frac{e(Bn-Di)}{1 + e(Bn-Di)}$ ; thereby expressing  $Pni$  as the log odds ratio of success.<sup>20,24,40</sup>

Thus, the probability that person  $n$  succeeds on item  $i$  is:  $Pni = \frac{e(Bn-Di)}{1 + e(Bn-Di)}$  with  $Pni$  expressed as log-odds units (logits);  $Bn$ =Ability person  $n$ ;  $Di$ =Difficulty item  $i$

For each item and person the expected score by the Rasch model can be calculated using the first formula (Table e-2). This is the dichotomous model as originally described by George Rasch.<sup>5</sup> However, since this first description several extended versions of the model have been described.<sup>26</sup>

**Table 1.** Raw scores

Items → / Persons ↓	a	b	c	d	e	f	g	h	i	Sum score	% (p) of success n/N (%)	Odds (p/ (1-p))	Ability (Log odds) (Bn)
Patient A	1	1	1	1	1	1	1	0	1	8	0.89 (89%)	8.09	2.09
B	1	1	1	1	1	0	1	1	0	7	0.78 (78%)	3.55	1.27
C	1	1	1	1	1	1	0	0	0	6	0.67 (67%)	2.03	0.71
D	1	0	1	1	1	1	0	0	0	5	0.56 (56%)	1.27	0.24
E	1	1	1	1	0	0	1	0	0	5	0.45 (45%)	0.82	-0.20
F	1	1	0	1	1	0	0	0	0	4	0.44 (44%)	0.79	-0.24
G	1	1	1	0	0	1	0	0	0	4	0.44 (44%)	0.79	-0.24
H	1	1	0	0	0	0	0	0	0	2	0.33 (33%)	0.49	-0.71
I	0	1	0	0	0	0	0	0	0	1	0.11 (11%)	0.12	-2.09
Sum score	8	8	6	6	5	4	3	1	1				
% (p) of success n/N (%)	0.89 (89%)	0.89 (89%)	0.67 (76%)	0.67 (76%)	0.56 (56%)	0.44 (44%)	0.33 (33%)	0.11 (11%)	0.11 (11%)				
Odds (p/(1-p))	8.09	8.09	2.03	2.03	1.27	0.79	0.49	0.12	0.12				
Difficulty item log odds (Di)	2.09	2.09	0.71	0.71	0.24	-0.24	-0.71	-2.09	-2.09				

**Table 2.** Expected values by the Rasch model

Difficulty item → / Ability patient ↓	2.09	2.09	0.71	0.71	0.24	-0.24	-0.71	-2.09	-2.09
2.09	0.50	0.50	0.80	0.80	0.86	0.91	0.94	0.98	0.98
1.27	0.31	0.31	0.64	0.64	0.74	0.82	0.88	0.97	0.97
0.71	0.20	0.20	0.50	0.50	0.62	0.72	0.81	0.94	0.94
0.24	0.14	0.14	0.38	0.38	0.50	0.62	0.72	0.91	0.91
-0.20	0.09	0.09	0.29	0.29	0.39	0.51	0.62	0.87	0.87
-0.24	0.09	0.09	0.28	0.28	0.38	0.50	0.62	0.86	0.86
-0.24	0.09	0.09	0.28	0.28	0.38	0.50	0.62	0.86	0.86
-0.71	0.06	0.06	0.19	0.19	0.28	0.38	0.50	0.80	0.80
-2.09	0.02	0.02	0.06	0.06	0.09	0.14	0.20	0.50	0.50



# **Defining a responder: its theoretical background**

## Introduction

To test the efficacy of a medical intervention within a patient population, the true score of interest should be measured (e.g. length or weight). For most qualities, the true score is unknown (e.g. for fatigue or disability). Surrogate outcome measures have been constructed and applied calculating an estimate score, which is considered a reflection of the true “unknown” score. However, imprecise calculations are always present for any estimated score and this is captured through the concept of Standard Errors (SE).<sup>1</sup> In fact, the estimated score is considered equivalent to the unknown true score + corresponding SE. In daily practice, for the assessment of clinical changes after an intervention, a fixed SE is assumed throughout the continuum of the scale being used. However, through the Rasch method, it has been demonstrated that each scale has a ‘U’-shape SE-pattern across the measure’s continuum (Figure 1), meaning that the SE varies across a scale’s range.

Researchers are able to use the SE values to calculate the confidence intervals around the unknown true score for the subsequent estimated scores. A standard error of  $\pm 1.96$  equals the interval around the true score within which we expect the estimate score to fall 95% of the time. A meaningful change over time is therefore considered if the estimated score  $\geq 1.96 \times \text{SE}$ : such a change is considered big enough to ascertain meaningfulness at a certainty of 95%.

## Responsiveness and effect size

Responsiveness (sensitivity to change) is defined as the ability of an instrument to detect important changes over time and is a required psychometric quality for any instrument to be used in clinical trials and research designs. Traditionally, sensitivity to change has been tested at the group-level by either comparing differences between two groups using p-values or by comparing the proportion of patients reaching an arbitrary pre-defined cut-off score on the metric being used.<sup>2,3</sup> The latter method compares e.g. improvement versus no-improvement based on a fixed SE assumption across the outcome measure used. Before improvement or no-improvement can be determined, the outcome measure of interest should be able to measure any changes over time between groups of patients for example by calculating the effect size. A higher effect size for the metric used indicates a higher discriminative ability between the two groups examined. However, a statistical significant difference between for example between intervention and placebo does not provide any information about the clinical relevance of such a difference, and even having high effect sizes can be misleading.<sup>4</sup>

Using modern techniques like the Rasch method, comparison at the ‘individual level

of being a responder' can be made using the model's provided individual standard errors for each patient at the various assessment points in time. Responsiveness at the individual level can report whether the obtained score of a particular patient has significantly improved or decreased. Meaningful change for groups of patients may not have the same significance for each individual within such a group.

## Minimum clinically important difference (MCID)

The clinical relevance of an effect can be determined by using the concept of a 'minimum clinically important difference' (MCID).<sup>5,6</sup> The MCID is considered "the smallest difference in score in the domain of interest that patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management. The MCID represents a change that would be considered meaningful and worthwhile by a patient such that they would consider repeating the intervention".<sup>6</sup> The MCID could enhance the interpretability of results of clinical studies and overcome the shortcomings of the 'statistically significant difference'.<sup>6-8</sup> As stated previously, a statistical significant difference between treatment and placebo does not always provide information about the clinical relevance. Also, obtaining statistical significance depends on the size of the trial and the outcome measure used.<sup>9</sup> Despite these disadvantages, clinical researchers often (mis)use the concept 'statistical significant difference'.<sup>10</sup> This may be attributed to preference of both researcher and journals for publishing positive results and the lack of knowledge about applying this concept.<sup>9</sup> The shortcomings of using only statistical significance have been recognized over the last years; these days several journals require reportage of both effect size and 95%-confidence intervals.<sup>9,11</sup> However, there are several methods (anchor-based and distribution-based) available to establish the MCID, and there is no consensus on the preferred method, which introduces another problem in the assessment of patients.<sup>7,8,12-14</sup> The biggest disadvantage of applying these MCID techniques is perhaps the standard error (SE) being wrongly considered "static (unchanged)" throughout the range of the outcome measure being used.

## Minimum Clinically Important Difference related to the varying Standard Errors (MCID-SE)

The Minimum Clinically Important Difference related to the varying Standard Error scores (MCID-SE) is based on the previously described significant change (SigChange) score.<sup>1,15</sup> For the individual patient, the level of change (through SigChange) can be determined (Figure 2). The MCID-SE can be computed by: a) their own individual change (personal-location at i month minus personal-location at entry), b) the

corresponding SE of difference related to their individual change ( $SE_{diff} = \text{square-root}(SE_{entry}^2 + SE_{i\text{ month}}^2)$ ), and c) the final MCID-SE calculations by dividing the individual change scores by corresponding  $SE_{diff}$  ( $MCID-SE = (\text{personal-location at } i \text{ month} - \text{personal-location at entry}) / SE_{diff}$ ).<sup>1</sup>

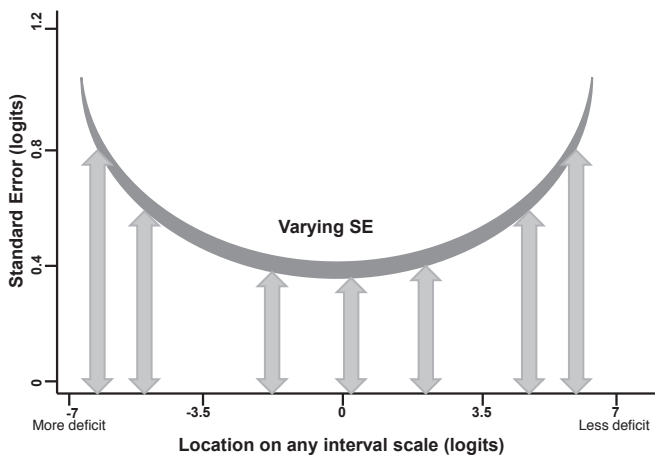
The degree of uncertainty associated with the ability estimate for each individual patient is expressed by the SE and is used to calculate the confidence intervals around the unknown true ability score. Thus, using 1 SE gives 68% certainty that the ability estimation is within this range; using 1.96 SE provides a 95% confidence.<sup>1,16</sup>

A meaningful change over time is generally considered if the estimated ability score  $\geq 1.96$  SE: this change is considered big enough to ascertain meaningfulness at a certainty of 95%. Therefore, based on the MCID-SE scores five different subgroups are created:

- subgroup 1 (clinically meaningful improvement):  $MCID-SE \geq 1.96$ ;
- subgroup 2 (clinically non-meaningful improvement):  $0 < MCID-SE < 1.96$ ;
- subgroup 3 (clinically unchanged):  $MCID-SE = 0$ ;
- subgroup 4 (clinically non-meaningful deterioration):  $-1.96 < MCID-SE < 0$ ;
- subgroup 5 (clinically meaningful deterioration):  $MCID-SE \leq -1.96$ .

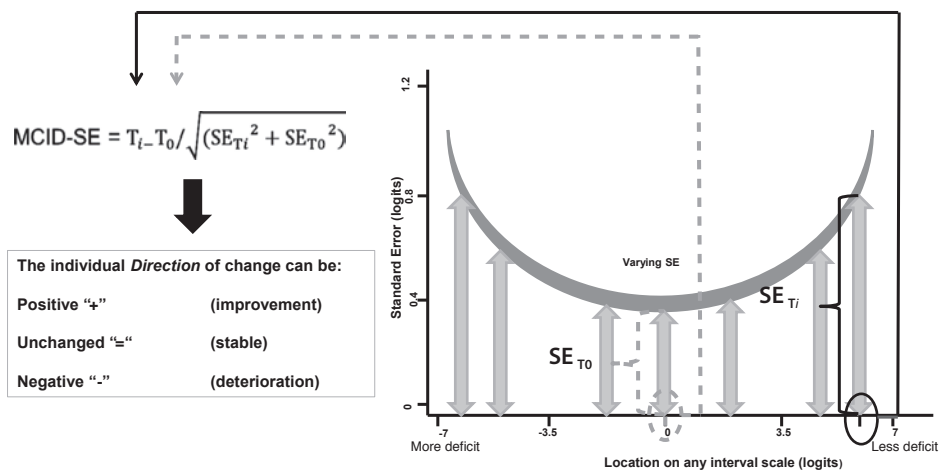
The definition of these subgroups (e.g. subgroup 1 showing clinically meaningful improvement) can be reversed depending on the construct of the scale being used.

**Figure 1.** Example of varying SE across the continuum of any interval scale



**Legend to Figure 1:** Schematic representation of the U-shape SE pattern across an outcome measure of interest. Modern methods like Rasch analysis have shown variable SE values related to the ability of a patient, with lower SE scores in the mid-sections of patient's ability and increasing SE values towards the extremes of patient's ability. SE=Standard Error.

**Figure 2.** Computing the Minimum clinically Important Difference related to the Standard Error score (MCID-SE)



**Legend to Figure 2:** The MCID-SE can be computed by dividing the individual change of a patient (personal-location at i month minus personal-location at entry) by the corresponding SE of difference related to their individual change ( $SE_{diff} = \text{square-root}(SE_{entry}^2 + SE_{1\text{ month}}^2)$ ).<sup>1</sup> MCID-SE = Minimum Clinically Important Difference related to the Standard Error scores. SE=Standard Error



## References

- 1 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health technology assessment* **13**, iii, ix-x, 1-177 (2009).
- 2 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet neurology* **7**, 136-144 (2008).
- 3 Leger, J. M., Chassande, B., Musset, L. *et al.* Intravenous immunoglobulin therapy in multifocal motor neuropathy: a double-blind, placebo-controlled study. *Brain* **124**, 145-153 (2001).
- 4 Hobart, J. C., Cano, S. J., and Thompson, A. J. Effect sizes can be misleading: is it time to change the way we measure change? *Journal of neurology, neurosurgery, and psychiatry* **81**, 1044-1048 (2010).
- 5 Nunnally, J. C. *Psychometric theory*. (McGraw Hill, 1978).
- 6 Jaeschke, R., Singer, J., and Guyatt, G. H. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* **10**, 407-415 (1989).
- 7 Sloan, J. A., Cella, D., Frost, M. *et al.* Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clin Proc* **77**, 367-370 (2002).
- 8 Merkies, I. S., van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *Journal of neurology, neurosurgery, and psychiatry* **81**, 1194-1199 (2010).
- 9 Sterne, J. A. and Davey Smith, G. Sifting the evidence-what's wrong with significance tests? *BMJ* **322**, 226-231 (2001).
- 10 Lesaffre, E. Use and misuse of the p-value. *Bulletin of the NYU hospital for joint diseases* **66**, 146-149 (2008).
- 11 Stang, A., Poole, C., and Kuss, O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol* **25**, 225-230 (2010).
- 12 Copay, A. G., Subach, B. R., Glassman, S. D. *et al.* Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* **7**, 541-546 (2007).
- 13 Norman, G. R., Sloan, J. A., and Wyrrwich, K. W. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* **41**, 582-592 (2003).
- 14 Beaton, D. E., Boers, M., and Wells, G. A. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol* **14**, 109-114 (2002).
- 15 Lai, J. and Eton, D. Clinically meaningful gaps. *Rasch Measurement Transactions* **14**, 850 (2002).
- 16 Wright, B. D. and Stone, M. H. *Best test design*. (Mesa Press, 1979).

## **Part II**

# **Impairment studies**



# **Improving Fatigue assessment in immune- mediated neuropathies: The modified Rasch-built Fatigue Severity Scale**

*Sonja I. van Nes<sup>1</sup>; Els K. Vanhoutte<sup>2</sup>; Catharina G. Faber<sup>2</sup>; Marcel Garssen<sup>3</sup>; Pieter A. van Doorn<sup>1</sup>;  
Ingemar S.J. Merkies<sup>1,2,4</sup>; on behalf of the PeriNomS study group*

<sup>1</sup>Department of Neurology, Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands;

<sup>2</sup>Department of Neurology, Maastricht University Medical Centre, Maastricht, the Netherlands;

<sup>3</sup>Department of Neurology, Jeroen Bosch Hospital, Den Bosch, the Netherlands;

<sup>4</sup>Department of Neurology, Spaarne Hospital, Hoofddorp, the Netherlands.

## Abstract

Fatigue is a major disabling complaint in patients with immune-mediated neuropathies (IN). The 9-item fatigue severity scale (FSS) has been used to assess fatigue in these conditions, despite having limitations due to its classic ordinal construct. The aim was to improve fatigue assessment in IN through evaluation of the FSS using a modern clinimetric approach [Rasch unidimensional measurement model (RUMM2020)]. Included were 192 stable patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP) or polyneuropathy associated with monoclonal gammopathy of undertermined significance (MGUSP). The obtained FSS data were exposed to RUMM2020 model to investigate whether this scale would meet its expectations. Also, reliability and validity studies were performed. The original FSS did not meet the Rasch model expectations, primarily based on two misfitting items, one of these also showing bias towards the factor 'walking independent.' After removing these two items and collapsing the original 7-point Likert options to 4-point response categories for the remaining items, we succeeded in constructing a 7-item Rasch-built scale that fulfilled all requirements of unidimensionality, linearity, and rating scale model. Good reliability and validity were also obtained for the modified FSS scale. In conclusion, a 7-items linearly weighted Rasch-built modified FSS is presented for more proper assessment of fatigue in future studies in patients with immune-mediated neuropathies.

## Introduction

Fatigue is considered a major disabling symptom in patients with immune-mediated neuropathies, such as Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and polyneuropathy associated with gammopathy of undetermined significance (MGUSP).<sup>1</sup> Despite an apparent good physical recovery, fatigue may lead to substantial decrement in quality of life expectations. The 9-item fatigue severity scale (FSS) has been used to capture the consequences of fatigue in patients with immune-mediated polyneuropathies and in many other chronic medical conditions. The FSS has demonstrated its simplicity, applicability, and various reliability and validity forms.<sup>2,3</sup> Parts of its construct validity were also based on factor analysis and its sum score has been recommended for scale score analyses. However, the use of a scale like the FSS has been criticized because it is based on the classical test theory (CTT), which recruits items without investigating their relevance for patients at examination.<sup>4</sup> Patients are requested to complete all the items in CTT, even though some may be inappropriate for their level of ability. Moreover, the suggested sum scores in CTT assume that given differences in these scores have equal meaning. Some caution is warranted when changes in sum scores are considered, since the whole score may not equal the sum of the parts.<sup>5</sup> Also, analyzing data using primarily factor analysis is prone to misleading results, because observations are non-linear which may generate illusory factors.<sup>6</sup>

More than 30 multi-item fatigue scales with good clinimetric properties have been developed and used for various chronic illnesses in the last decades.<sup>7</sup> However, a linearly weighted fatigue scale is preferred using modern scientific methods in patients with immune-mediated neuropathies. The Rasch technology is increasingly being recognized as a modern test method that could facilitate in this.<sup>8-10</sup> This technology attempts to transform ordinal scores into linear measures and is based on a logical assumption: individuals with high ability to perform a task should have an increased probability, relative to individuals with low levels, of getting a better score on any item.<sup>11</sup>

The aim of this study is to improve fatigue assessment in patients with immune-mediated neuropathies. Therefore, we examined whether the original FSS scale fulfils model expectations using a Rasch measurement model. Deviation from model expectation will be examined to determine whether the FSS construct can be improved. Finally, reliability and validity aspects will be investigated.

## Patients and methods

### Patients

Data obtained from 192 patients (163 with GBS, 21 with CIDP, and 8 with MGUSP) were re-examined for the purposes of this study. Patients were recruited from the Erasmus Medical Center Rotterdam databank and the Dutch GBS study group and have already contributed to the clinimetric studies of the INCAT group and a published randomized trial evaluating fatigue in GBS.<sup>3,12</sup> The following inclusion criteria were applied: a stable clinical condition was mandatory and was defined as an unchanged physical or functional condition in the last 6 months before the start of the studies as declared by the patients to the best of their knowledge. Where possible, findings at physical examination at entry were compared with available earlier physical examination data to ascertain stability. Patients were excluded from participation if there was any concomitant disease or use of medication that might cause chronic fatigue. All selected patients declared to have experienced only mild and transient fatigue prior to their illness and still had residual symptoms or signs resulting from their illness (f-score  $\geq 1$ ).<sup>13</sup> The patients with GBS and CIDP met the international criteria for their illness.<sup>14,15</sup> The diagnosis MGUSP was established after excluding other underlying causes for the gammopathy and polyneuropathy.<sup>16</sup>

### Assessment scales

The FSS, a brief and simple self-assessed questionnaire, has demonstrated its scientific soundness in immune-mediated neuropathies.<sup>1,2</sup> The FSS is a 9-items questionnaire with response categories ranging from 1 “strongly disagree”, 2 “mainly disagree”, 3 “partially disagree”, 4 “do not agree/disagree”, 5 “partially agree”, 6 “mainly agree”, to 7 “strongly agree” for each inquiry. The sum score of the nine inquiries ranges from 9 (“no signs of fatigue”) to 63 (“most disabling fatigue”).

The *short-form fatigue scale (SFFS)*, with acceptable validity and reliability, was administered in order to evaluate the construct convergent validity of the analysed FSS. The score of this scale ranges from 4 (“no signs of fatigue”) to a maximum of 28 points (“most severe fatigue”).<sup>1,17</sup>

### Test procedure

FSS data were obtained with informed written consent and after approval of the medical ethical committee of the Erasmus Medical Center. Participants were lucid and received instructions on how to complete the fatigue forms. These questionnaires were answered in random order. Patients were examined at our outpatient clinic. Scores were double-checked before departure of the patients. The FSS was mailed 2-4 weeks later to all patients (2<sup>nd</sup> sample) for a second assessment (test-retest).

## General Statistics and Rasch measurement aspects

### • Descriptive statistics

Personal aspects like age, gender, type of inflammatory neuropathy, duration of illness, and GBS disability level (f-score) were collected. For Rasch analyses purposes, these personal factors were arbitrarily categorised as follows: age category (1: < 40 years, 2: 40-59 years, 3: ≥ 60 years), gender (0: female, 1: male), illness type (1: GBS; 2: CIDP; 3: MGUSP), polyneuropathy form (1: acute = GBS; 2: chronic = CIDP + MGUSP patients), duration category (1: < 2 years; 2: 2-5 years, 3: ≥ 5 years), and walking independent category (0: unable = f-score > 2; 1: able = f-score ≤ 2). The rationale for these categories was to investigate their possible influence on experiencing fatigue.<sup>18</sup>

### • Sample size consideration

For Rasch analyses purposes, a convenient sample size of 150 patients has been suggested to estimate item difficulty, with a confidence interval of 99% and item calibrations within +0.5 logits.<sup>19</sup> In the current study, a sample of 192 patients was available, thus expecting an acceptable degree of precision of the Rasch analysis.

### • Rasch measurement model

The Rasch model is based on the probabilistic Guttman procedure.<sup>20</sup> For the purposes of the current study, the obtained FSS data and the selected 'personal factors' were subjected to Rasch analysis using the RUMM2020 software.<sup>11,21,22</sup>

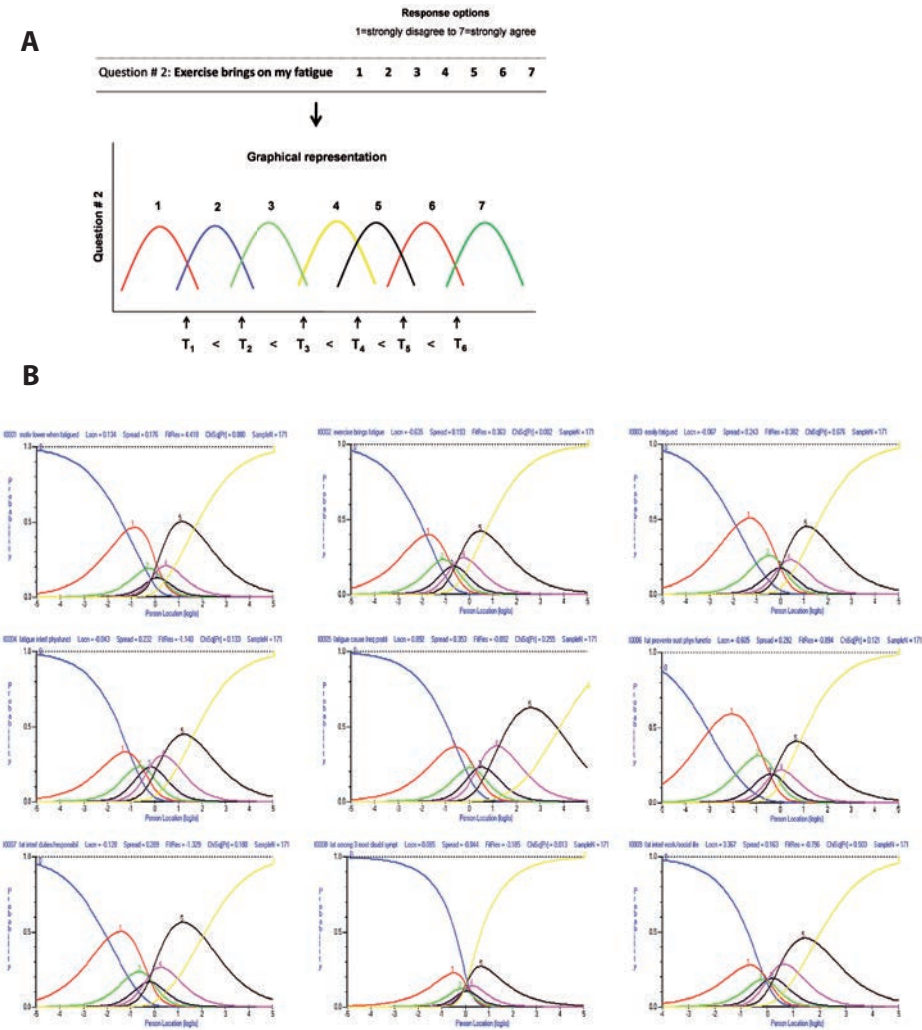
The following applications of RUMM2020 were used to examine whether the data fit the model expectations and items or patients not fulfilling these requirements were removed or subjected to re-adjustments to fit the model:

- Fit statistics: Three overall fit statistics are considered. Two are item-person interaction statistics transformed to approximate a z score. Therefore, if the items and persons fit the model, a mean around zero and a standard deviation of 1 would be expected. A third fit statistic score is an item-trait interaction statistic reported as a chi-square, reflecting the property of invariance across the trait. A non-significant chi-square indicates that the hierarchical ordering of the items does not vary across the trait, thus fulfilling the required property of invariance.
- Internal reliability studies: Internal consistency reliability of the scale was determined with the Person Separation Index (PSI) or Cronhbach's alpha (the latter is only possible in case of no missing values) using the logit scores for each person. A value of ≥ 0.7 was considered consistent with the scale being able to differentiate between at least 2 groups of patients and is seen as the minimum requirement for measurement.<sup>23</sup>
- Fit residuals: Individual person- and item-fit statistics were examined, both as residuals (deviations of items and persons from expected model scores: residuals between ±2.5 are considered adequate fit) and using a chi-squared statistic (significance indicating misfit).



- Threshold examination: The term threshold refers to the point between two adjacent response categories where either response is equally probable. That is the point where, for example, in case of the FSS, a threshold would be the point between two adjacent response categories for each item (e.g., between 1: strongly disagree and 2: mainly disagree; see Figure 1A). One of the most common sources of item misfit concerns respondents' inconsistent use of these response options. This results in what is known as 'disordered thresholds': the failure of respondents to use the response categories in a manner consistent with the level of the trait being measured. If needed, response options were collapsed in order to improve overall fit to the model.
  - Local dependency: Residual correlations between items within the same scale are a source of misfit. This local dependency occurs when a respondents' answer to one item automatically has a bearing upon the answer to another item. Local dependency was examined, since it affects the estimation of test information and item discrimination parameters, thereby inflating the final scale score in a particular direction.<sup>24</sup>
  - Item bias: A scale should always work in the same way irrespective of which personal factor (e.g., gender) is being examined.<sup>25</sup> For example, if men and women with equal levels of disability state for their inflammatory neuropathy respond systematically differently to a fatigue item, then this item would be considered to display differential item functioning (DIF = item bias), and would violate the requirement for unidimensionality. The obtained data were examined for DIF using statistical (analyses-of-variance) and graphical methods.
  - Model forms: The unrestricted partial credit Rasch model was used in the current study. Finally, the Likelihood Ratio Test was applied to examine whether the final obtained model would fulfil the Rating scale model requirements.
  - Test for unidimensionality: Unidimensionality is tested by allowing the factor loadings on the first residual to determine subsets of items and then testing (paired *t* test) to see if the person estimate (logit of person 'ability' or, in this case "degree of fatigue") derived from these subsets significantly differs from that obtained from all items. The absence of any meaningful pattern in the residuals will support the assumption of local independence and unidimensionality of the scale.<sup>26</sup>
- **External construct validity and reliability studies**  
The external construct validity of the final modified FSS scale was assessed through convergent validity with the SFFS (intraclass correlation coefficient reported).<sup>17</sup> Also, as part of the final analyses, graphical test-retest validity and reliability studies were performed comparing items' hierarchy and patients' location between the two samples of patients.<sup>27</sup> The obtained correlations were quantified by estimation of the intraclass correlation coefficient using a one-way random effects analysis-of-variance model for groups' comparison.

Figure 1. Category probability curves



**Legend to Figure 1:** A: A graphical representation of the ideal ability of patients to discriminate between response options of fictitious items ranging from 1 (strongly disagree) to 7 (strongly agree). B: Patients with immune-mediated neuropathies demonstrated unanimously inability to differentiate between the original response options of the original 9-item fatigue severity scale (FSS) ranging from 2 ("mainly disagree") to 6 ("mainly agree"), leading to many inverted thresholds.

- **Software**

RUMM2020 Rasch software was used and further statistical and graphical analyses were undertaken using Stata 10.0 for Windows XP (Stata Statistical Software: Release 10.0, TX: Stata Corporation 2007). A value of  $p = 0.0055$  ( $0.05/9$ ) was used throughout the analyses, based on Bonferroni multiple testing corrections.<sup>28</sup>

## Results

### General description of patients

A total of 192 patients with immune-mediated neuropathies have been included in this study. The basic characteristics of all patients are presented in Table 1. A total of 183 patients returned the second FSS assessment (95.3%). Forty-five patients (23.4%) were not able to walk independently (f-score > 2).

**Table 1.** Basic characteristics of patients with inflammatory polyneuropathies

	First sample	Second sample (for test-retest)
Number of patients	192	183
Age, mean years (SD), range	52.9 (14.3), 17-84	52.3 (14.1), 17-84
Gender, # (%)		
Females	94 (49)	91 (49.7)
Males	98 (51)	92 (50.3)
GBS; # (%)	163 (84.9)	158 (86.3)
CIDP; # (%)	21 (10.9)	20 (10.9)
MGUSP; # (%)	8 (4.2)	5 (2.7)
FSS mean (SD), range	52.1 (10.3), 9-63	50.2 (11.2), 9-63
GBS disability score (f-score); # (%)		
1	50 (26.0)	46 (25.1)
2	97 (50.5)	95 (51.9)
3	34 (17.7)	31 (16.9)
4	11 (5.7)	11 (6.0)
Duration of symptoms, years, Mean (SD), range	5.5 (5.9), 0-41	5.2 (5.0), 0-28

**Legend to Table 1:** GBS = Guillain-Barré syndrome; CIDP= chronic inflammatory demyelinating polyradiculoneuropathy; MGUSP = monoclonal gammopathy of undetermined significant related polyneuropathy; FSS = fatigue severity scale.

### Initial Rasch studies on the FSS

#### • Fit statistic description

The original 9-items FSS scale did not meet the Rasch model expectations. There was a substantial items misfit with a mean fit residual of -0.253 and a SD of 2.058. The patients demonstrated reasonable fit statistics (mean fit residual -0.350, SD 1.025). There was a significant item-trait chi-square probability ( $p=0.000006$ ) indicating a highly significant deviation from the model. The Cronbach's alpha was 0.92 and the PSI 0.93. Individual item fit statistics were also investigated and presented in Table 2. As can be seen, items 1 and 8 demonstrated a misfit due to fit residuals exceeding  $\pm 2.5$  (item 1: +4.418 and item 8: -3.185). Item 1 had also a significant low chi-square probability. The high negative fit residual of item 8 demonstrates its redundancy to the data set. Item 8 is not adding any information to the data.

**Table 2.** Rasch fit statistics for the initial 9-item fatigue severity scale (FSS) data

FSS item #	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	0.134	0.07	4.418	148.11	26.828	2	0.000001
2	-0.635	0.096	0.363	148.11	5.002	2	0.081988
3	-0.067	0.073	0.382	148.11	0.784	2	0.675829
4	-0.043	0.078	-1.14	148.11	4.032	2	0.133195
5	0.892	0.068	-0.092	148.11	2.736	2	0.254641
6	-0.605	0.085	-0.894	148.11	4.229	2	0.120719
7	-0.128	0.082	-1.329	148.11	3.424	2	0.1805
8	0.085	0.069	-3.185	148.11	8.66	2	0.013165
9	0.367	0.068	-0.796	148.11	1.376	2	0.50254

**Legend to Table 2:** Items 1 and 8 demonstrated a misfit due to fit residuals exceeding  $\pm 2.5$ . Item 1 had also a significant low chi-square probability. The high positive labelled fit residual and significant probability of item 1 suggests that the response to this item differed from the responses to the remaining scale items. The high negative fit residual of item 8 demonstrates its redundancy to the data set; in other words, item 8 is not adding any information to the data. SE= standard error, FitResid = Fit residuals, DF = degrees of freedom, ChiSq= Chi square, prob = probability.

- **Threshold examination (category probability curves analyses)**

All items demonstrated in the mid response category area (response options ranging from 2: “mainly disagree” up to 6: “mainly agree”) a general inability of the patients to differentiate between these options, with disordered thresholds for almost all items. Item 8 (“Fatigue is among my three most disabling symptoms”) demonstrated the strongest disordered threshold pattern (Figure 1B).

- **Local dependency (item residual correlations) and Item bias analyses**

No significant correlations were seen between the original nine items of the FSS. All items were subjected to DIF investigation for the selected personal factors. Only for item 8, a uniform DIF was demonstrated ( $p=0.000562$ ; see Figure 2). Item 8 was considered more difficult to answer for the 45 patients who could not walk independently. The remaining items did not show any item bias.

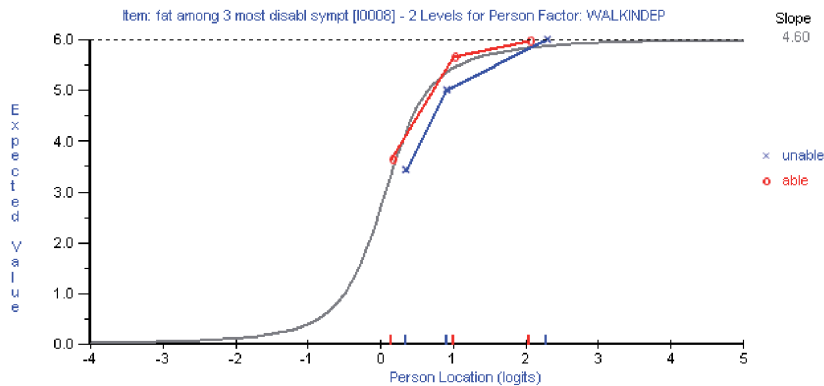
- **Independent t-test**

Based on the first principal components analysis, two item subsets were formed by grouping the three most positive loading items (item 1: loading 0.685; item 2: 0.661; item 6: 0.190) vs. the three most negative loading items (item 5: loading -0.557; item 9: -0.433; item 7: -0.390). The two subsets were compared through independent t-test and demonstrated a proportion of 0.0819, suggesting that the data did not show unidimensionality.

### **Data handling to improve the FSS fit to the Rasch model**

- **Rescoring the items to improve thresholds and removing items**

Based on the findings illustrated in Figure 1, we decided to systematically (from item 1 through 9) rescore the items in order to restore reversed thresholds. Items 1, 2, 4, 5, 7, 8, and 9 were rescored as follows: original scale scores 1-7, initial Rasch scores 0 - 6;

**Figure 2.** Item characteristic curve for item 8 related to personal factor 'ability to walk independent.'

**Legend to Figure 2:** Patients unable to walk independently demonstrated uniform differential item functioning (DIF) on item 8: this item was experienced as more difficult to answer compared with those able to walk independent.

rescoring: 0 = 0, 1-3 = 1; 4-5 = 2, and 6 = 3. Items 3 and 6 were rescored as follows after studying the location of disordered thresholds: 0 = 0, 1-2 = 1; 3-5 = 2, and 6 = 3. After performing these, items 1 and 8 still demonstrated misfit statistics and were subsequently removed from the analyses.

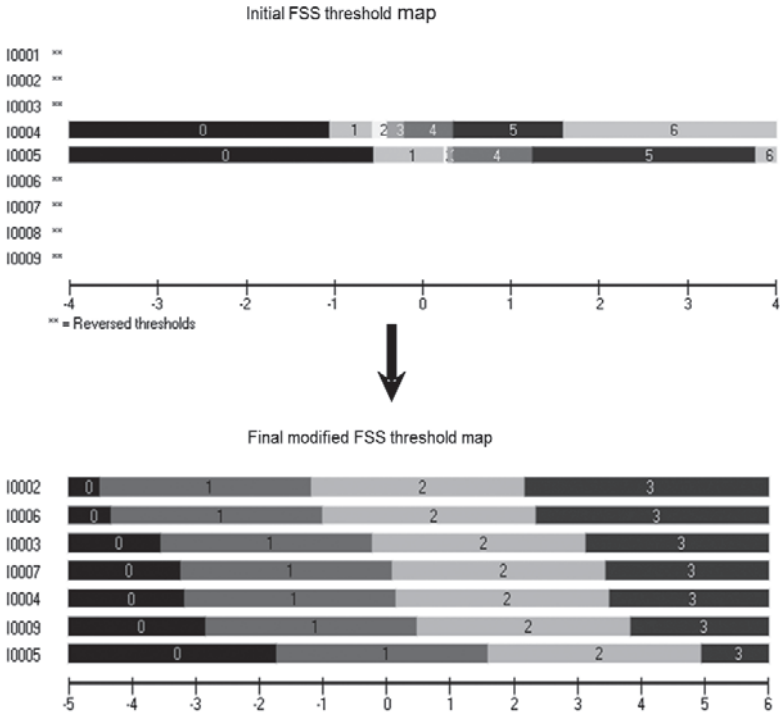
- Threshold mapping, fit statistics and testing for Rating scale model expectations**  
 With the help of threshold mapping, we were able to restore the disordered thresholds and transformed the response categories in the remaining seven items to equivalent 4-points response categories (Figure 3 and Appendix). For the items and patients, mean fit residuals of -0.289 (SD: 1.113) and -0.275 (SD: 1.091) were obtained, respectively. The Cronbach's alpha and PSI remained good (0.89 for both) with an overall chi-square probability of 0.22. All individual items had good fit statistics (Table 3). Item 2 ("exercise brings on my fatigue") turned out to be the easiest inquiry to answer while item 5 ("fatigue causes frequent problems for me") was considered the most difficult (Table 3). A p-value of 0.10 was obtained for the modified 7-items FSS version by means of a Likelihood Ratio test, implicating that the items meet the Rasch Rating scale model expectations.
- Item bias and test for unidimensionality**  
 The individual items were all free from DIF. Based on the first principal components analysis, two item subsets were formed again: positive loading residual items (2, 3, and 6) vs. negative items (5, 7, and 9). The independent *t* test between these two subsets of items demonstrated a proportion of 0.02, which shows unidimensionality. Finally, the logits for the patients, ranging from -6 to +6, were transformed into a raw 7-item FSS sum score, ranging from 7 ("no signs of fatigue") to 21 ("most disabling fatigue") using the equation " $y = 17.42 + (1.73 \times \text{person location [logits]})$ ". Only 1.1% and 9.7% of the patients sample size had a floor or ceiling effect, respectively.

**Table 3.** Rasch fit statistics for the final 7-item fatigue severity scale (FSS) data

FSS item #	Location	SE	FitResid	DF	ChiSq	DF	Prob
2	-1.174	0.165	1.95	140.43	0.109	2	0.947176
6	-0.995	0.162	-1.271	140.43	3.082	2	0.2142
3	-0.208	0.153	-0.874	140.43	1.137	2	0.566385
7	0.11	0.15	-1.139	140.43	2.83	2	0.242876
4	0.158	0.15	-0.197	140.43	3.311	2	0.19102
9	0.502	0.148	0.721	140.43	4.186	2	0.12334
5	1.607	0.145	-1.16	140.43	2.991	2	0.224179

**Legend Table 3:** Items 1 and 8 were removed due to misfitting the model and the remaining 7 items met the model requirements. Loading on location indicates item difficulty in patients with immune-mediated neuropathies (most negative = easiest, most positive = most difficult item to answer) SE= standard error, FitResid = Fit residuals, DF = degrees of freedom, ChiSq= Chi square, prob = probability.

**Figure 3.** Rescoring response categories



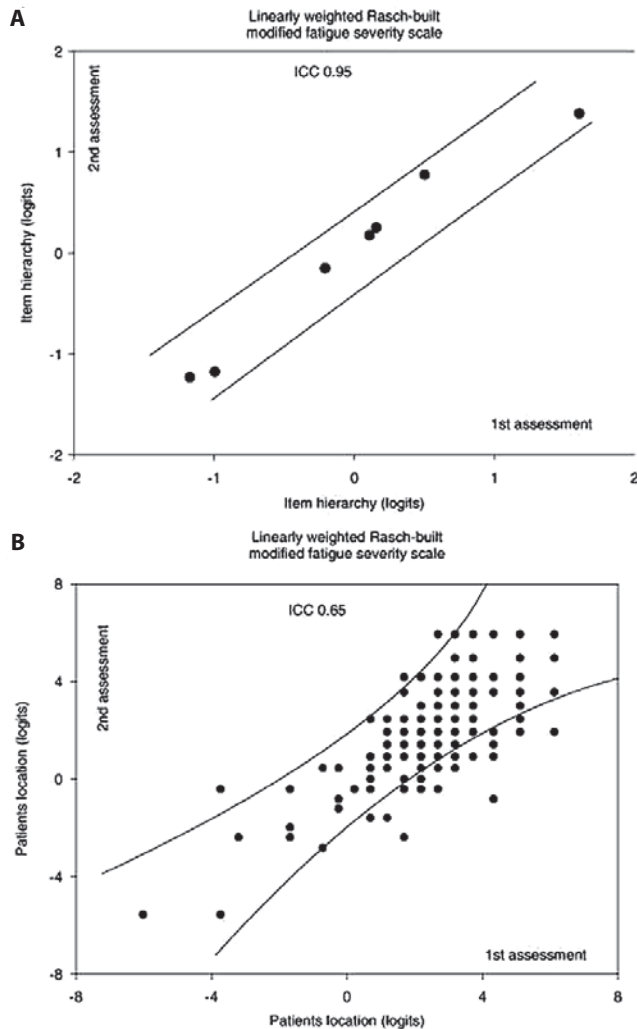
**Legend to Figure 3:** After deleting misfitting items and rescoring the response options (from seven categories to four categories), we were able to construct a modified FSS 7-items scale fulfilling all Rasch model expectations.

### External construct validity and reliability studies

The modified FSS demonstrated an acceptable correlation with the SFFS (intraclass correlation coefficient: 0.71,  $p < 0.0001$ ). Test retest reliability studies were performed

to investigate whether the items' hierarchy and patients' location were consistent in time. As demonstrated in Figure. 4, all 7 items were located within the 95% CI lines, indicating ideal invariance. Most patients were also within the 95% CI limits.

**Figure 4.** Test-retest reliability studies between item hierarchy (A) and person location (B) for the modified fatigue severity scale (FSS) in immune-mediated neuropathies.



**Legend to Figure 4:** A: Test-retest reliability studies between items' hierarchy assessed at the first and second FSS assessment and the 95% of confidence interval (solid lines) for the ideal invariance. As can be seen, all items were ordered in a hierarchal way, thus demonstrating ideal invariance (intraclass correlation coefficient for the item estimates of 0.95). B: Test-retest reliability studies between persons' location assessed at the first and second FSS assessment and the 95% of confidence interval (solid lines) for the ideal invariance. Acceptable reproducibility was obtained between the patients' locations (intraclass correlation coefficient for the item estimates of 0.75).

## Discussion

To our knowledge, the current study is the first study examining the modern clinimetric aspects for the FSS in patients with immune-mediate neuropathies. The original 9-item FSS failed in meeting the requirements addressed by the Rasch model.<sup>9</sup> However, after systematic evaluation and re-adjustment of the data, we were able to present a linearly weighted 7-item (4 response categories for each item) Rasch-built modified FSS, that fulfilled all clinimetric requirements, including validity and reliability (see also Appendix).<sup>10,11,27</sup> The patients' location was acceptable, although most patients were within the 95% confidence interval limits and having high scores which indicates having a more impaired status due to severe fatigue (Figure 4B). This is in conformity with earlier findings.<sup>1</sup> An elegant and equivalent study on FSS has been recently presented, which demonstrated quite similar findings in patients with MS.<sup>29</sup> In this paper, more items were disordered and more DIF was seen in patients at examination. The final modified FSS in patients with MS turned out to be somewhat different. It is conceivable that the obtained differences in experiencing fatigue between patients with MS vs. those in our study could be related to differences in possible confounding factors such as location of illnesses (central vs. peripheral nervous system), differences in disability, and possible differences in pathophysiological mechanism of fatigue in these disorders.

In the current study, the most strikingly pattern was the patients' unanimous inability to differentiate between the original response categories ranging from 2 to 6, leading to many inverted thresholds (Figure 1 and 3). Although the authors of the original FSS attempted to improve differentiation between patients, the amount of response categories led to confusion in patients with immune-mediated neuropathies that could only be visualized using a modern technique like the Rasch. This was also in conformity with earlier reports demonstrating difficulties of patients to discriminate among more than 3 response categories.<sup>30</sup>

The obtained 7-item Rasch-built modified FSS has the advantages of a modern test theory (e.g., linearity, unidimensionality) and its use is believed to improve fatigue assessment in immune-mediated neuropathies and will certainly help clinicians to focus on ameliorating the enormous disabling impact of fatigue. Therefore, we suggest the use of this modified 7-items FSS in future studies targeting fatigue in patients with GBS, CIDP, and MGUSP.



Appendix

Transforming 9- item FSS to 7-item FSS

Original Fatigue severity scale								
1=strongly disagree; 2=mainly disagree; 3=partially disagree; 4=do not agree/disagree; 5=partially agree; 6=mainly agree 7=strongly agree (circle one answer per question)								
1	My motivation is lower when I am fatigued	1	2	3	4	5	6	7
2	Exercise brings on my fatigue	1	2	3	4	5	6	7
3	I am easily fatigued	1	2	3	4	5	6	7
4	Fatigue interferes with my physical functioning	1	2	3	4	5	6	7
5	Fatigue causes frequent problems for me	1	2	3	4	5	6	7
6	My fatigue prevents sustained physical functioning	1	2	3	4	5	6	7
7	Fatigue interferes with carrying out certain duties and responsibilities	1	2	3	4	5	6	7
8	Fatigue is among my three most disabling symptoms	1	2	3	4	5	6	7
9	Fatigue interferes with my work, family, or social life	1	2	3	4	5	6	7



Rasch-built 7-items modified Fatigue severity scale				
The higher the score you choose, the more you agree with the question (the lower the score the less you agree )				
	Disagree			Agree
1	Exercise brings on my fatigue	1	2	3 4
2	I am easily fatigued	1	2	3 4
3	Fatigue interferes with my physical functioning	1	2	3 4
4	Fatigue causes frequent problems for me	1	2	3 4
5	My fatigue prevents sustained physical functioning	1	2	3 4
6	Fatigue interferes with carrying out certain duties and responsibilities	1	2	3 4
7	Fatigue interferes with my work, family, or social life	1	2	3 4

Score range: 7 ("no signs of fatigue") to 21 ("most disabling fatigue").

## References

- 1 Merkies, I. S., Schmitz, P. I., Samijn, J. P. *et al.* Fatigue in immune-mediated polyneuropathies. European Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **53**, 1648-1654 (1999).
- 2 Krupp, L. B., LaRocca, N. G., Muir-Nash, J. *et al.* The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch Neurol* **46**, 1121-1123 (1989).
- 3 Merkies, I. S. J. *Evaluation of scales and measurement instruments in immune-mediated polyneuropathies. Thesis*, Erasmus Medical Center, (2001).
- 4 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-59 (2006).
- 5 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* **49**, 711-717 (1996).
- 6 Linacre, J. M. Rasch First or Factor First? 11:4 p. 603. *Rasch Measure Trans* **11**, 603 (1998).
- 7 Dittner, A. J., Wessely, S. C., and Brown, R. G. The assessment of fatigue: a practical guide for clinicians and researchers. *J Psychosom Res* **56**, 157-170 (2004).
- 8 Bond, T. G. and Fox, C. M. *Applying the Rasch model: fundamental measurement for the human sciences.* (Lawrence Erlbaum Associates, 2001).
- 9 Rasch, G. *Probabilistic models for some intelligence and attainments tests.* (1980).
- 10 Streiner, D. L. and Norman, G. R. *Health measurement scales. A practical guide to their development and use.* (Oxford University Press, 2nd ed, 1998).
- 11 Pallant, J. F. and Tennant, A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* **46**, 1-18 (2007).
- 12 Garssen, M. P., Schmitz, P. I., Merkies, I. S. *et al.* Amantadine for treatment of fatigue in Guillain-Barre syndrome: a randomised, double blind, placebo controlled, crossover trial. *J Neurol Neurosurg Psychiatry* **77**, 61-65 (2006).
- 13 Kleyweg, R. P., van der Meche, F. G., and Schmitz, P. I. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle Nerve* **14**, 1103-1109 (1991).
- 14 Asbury, A. K. and Cornblath, D. R. Assessment of current diagnostic criteria for Guillain-Barre syndrome. *Ann Neurol* **27 Suppl**, S21-24 (1990).
- 15 AANA. Research criteria for diagnosis of chronic inflammatory demyelinating polyneuropathy (CIDP). Report from an Ad Hoc Subcommittee of the American Academy of Neurology AIDS Task Force. *Neurology* **41**, 617-618 (1991).
- 16 Miescher, G. C. and Steck, A. J. Paraproteinaemic neuropathies. *Baillieres Clin Neurol* **5**, 219-232 (1996).
- 17 Alberts, M., Smets, E. M., Vercoulen, J. H. *et al.* [Abbreviated fatigue questionnaire: a practical tool in the classification of fatigue]. *Ned Tijdschr Geneeskde* **141**, 1526-1530 (1997).
- 18 Jonker, A. A., Comijs, H. C., Knipscheer, K. C. *et al.* Persistent Deterioration of Functioning (PDF) and change in well-being in older persons. *Aging Clin Exp Res* **20**, 461-468 (2008).
- 19 Linacre, J. M. Sample size and item calibration stability. *Rasch Measure Trans* **7**, 28 (1994).
- 20 Altman, D. G. *Practical statistics for medical research.* (1990).
- 21 Andrich, D., Lyne, A., Sheridan, B. *et al.* *RUMM 2020.* (2003).
- 22 Smith, R. M. Fit analysis in latent trait measurement models. *J Appl Meas* **1**, 199-218 (2000).
- 23 Fisher, W. P. Reliability Statistics. *Rasch Measure Trans* **6**, 238 (1992).
- 24 Embretson, S. E. and Reise, S. P. *Item response theory for psychologists.*, (2000).
- 25 Dorans, N. J. and Holland, P. W. in *Differential item functioning* (eds P. W. Holland & H. Wainer) 36-66 (Lawrence Erlbaum Associates, 1993).
- 26 Smith, E. V., Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* **3**, 205-231 (2002).
- 27 Wright, B. D. and Stone, M. H. *Best test design.* (Mesa Press, 1979).
- 28 Bland, J. M. and Altman, D. G. Multiple significance tests: the Bonferroni method. *BMJ* **310**, 170 (1995).
- 29 Mills, R., Young, C., Nicholas, R. *et al.* Rasch analysis of the Fatigue Severity Scale in multiple sclerosis. *Mult Scler* **15**, 81-87 (2009).
- 30 Vandervelde, L. *Activity limitations in patients with neuromuscular disorders*, Univesite catholique de Louvain, (2008).



# Vigorimeter Grip strength in CIDP: a responsive tool that rapidly measures the effect of IVIG- the ICE study

*Els K. Vanhoutte<sup>1</sup>; Norman Latov<sup>2</sup>; Chunqin Deng<sup>3</sup>; Kim Hanna<sup>3</sup>; Richard A. C. Hughes<sup>4</sup>; Vera Bril<sup>5</sup>; Marinos C. Dalakas<sup>6</sup>; Peter Donofrio<sup>7</sup>; Pieter A. van Doorn<sup>8</sup>; Hans-Peter Hartung<sup>9</sup>; Ingemar S. J. Merkies<sup>10</sup>*

<sup>1</sup>Department of Neurology, Maastricht University Medical Centre, Maastricht, the Netherlands;

<sup>2</sup>Peripheral Neuropathy Center, Cornell University, New York, NY, USA; <sup>3</sup>Grifols Inc., Research Triangle Park, NC, USA; <sup>4</sup>MRC Centre for Neuromuscular Disease, National Hospital for Neurology and Neurosurgery, Queen Square, London, UK; <sup>5</sup>University Health Network and University of Toronto, Toronto, Ontario, Canada; <sup>6</sup>Department of Neurology, Thomas Jefferson University, Philadelphia, PA, USA; <sup>7</sup>Department of Neurology, Vanderbilt University, Nashville, TN, USA; <sup>8</sup>Department of Neurology, Erasmus MC, Rotterdam, the Netherlands; <sup>9</sup>Department of Neurology, Heinrich Heine University, Düsseldorf, Germany; <sup>10</sup>Department of Neurology, Spaarne Hospital, Hoofddorp, the Netherlands

## Abstract

In a recent trial in patients with chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), the ICE study, grip strength measurement captured significantly more improvement in patients receiving intravenous immune globulin (IGIV-C) than in those receiving placebo.

We conducted a systematic analysis to determine the sensitivity of grip strength as an indicator of meaningful clinical changes in CIDP.

A randomized double-blind trial was undertaken in 117 CIDP patients who received IGIV-C or placebo every 3 weeks for up to 24 weeks. Grip strength and inflammatory neuropathy cause and treatment (INCAT) disability scores were assessed at each visit, and the responsiveness of each scale was compared. A minimum clinically important difference cut-off value for grip strength ( $> 8$  kPa) and INCAT score ( $> 1$  point) was applied to assess the proportion of responders to IGIV-C versus placebo. This analysis showed that grip strength demonstrated significant improvement earlier (as early as day 16) than the INCAT disability scale in patients receiving IGIV-C compared with placebo. A significantly higher proportion of improvers were seen in the IGIV-C group (37.5–50.9%) than the placebo group (21.1–25.9%) for grip strength at day 16, week 3, week 6, and the end of the first period. Also, grip strength showed within the first 6 weeks in the placebo group significantly more patients with a clinically meaningful deterioration ( $> 8$  kPa), compared to the INCAT ( $> 1$ -point deterioration) findings.

Grip strength can be considered a sensitive tool for assessing clinically relevant changes in patients with CIDP. Its use in daily practice is suggested.

## Introduction

At a consensus meeting on outcome measures in inflammatory neuropathies, the level of disability was chosen to be the primary measure of assessing efficacy in these patients.<sup>1</sup> Therefore, the adjusted inflammatory neuropathy cause and treatment (INCAT) disability score – a 10-point ordinal measure capturing problems in daily arm and leg activities and mobility – was chosen as the primary outcome in the largest trial ever performed in patients with chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), the ICE study.<sup>2,3</sup>

This disability measure has demonstrated statistically significant differences in favour of those patients being treated with intravenous human immune globulin, 10% caprylate/chromatography purified (IGIV-C) compared with those receiving placebo.<sup>3</sup> However, while the adjusted INCAT disability score is a very useful tool with which to judge treatment response, it is not the typical measurement employed by neurologists in daily practice. By contrast, grip strength as measured with the Vigorimeter, which is a simple tool that is easy to apply and provides an instantaneous result.<sup>4</sup> Grip strength using a hand-held dynamometry or an equivalent air-pressure tool as the Martin Vigorimeter, is relatively objective and less susceptible to bias, but it is limited as an outcome measure because impairment is assessed solely for the hands.<sup>5</sup> In the ICE trial, grip strength also demonstrated a statistically significant improvement in favour of IGIV-C over placebo.<sup>3</sup> However, a systematic investigation has not been performed determining the sensitivity of grip strength versus the INCAT disability score as indicators of relevant clinical changes in CIDP. The purpose of the study reported here was to systematically examine the sensitivity of grip strength measurement and establish whether changes measured by this tool during a period of IGIV-C administration versus placebo are clinically meaningful and thus indicative of a tool that could be used also in daily practice as a good indicator of improvement or worsening in patients with CIDP.

## Methods

### Study design

For the purposes of the current study, the ICE trial first period data for the INCAT disability score and dominant hand grip strength scores were used.<sup>3</sup> Complete details of the patient population, study design, and methodology are described in the publication of the primary efficacy analysis.<sup>3</sup> The study design (ClinicalTrials.gov identifier: NCT00220740) was approved by a regional review board. Briefly, the ICE study was a randomized, double-blind, placebo-controlled, response-conditional (rescue) trial and included a first period and a crossover (rescue) period.<sup>3</sup> Eligible participants were randomized to receive either IGIV-C (Gamunex®, Grifols Therapeutics,

Inc., formally Talecris Biotherapeutics Inc., Research Triangle Park, NC) or placebo (0.1% albumin). Participants received a baseline loading dose of 2 g/kg (commencement of first or crossover periods) and then a maintenance infusion of 1 g/kg every 3 weeks for up to 24 weeks. An adjusted INCAT disability score was applied to assess functional disability prior to every subsequent infusion of study drug. Any participant with an adjusted INCAT disability score that worsened from baseline by  $\geq 1$  point at any time from day 16 until week 24 during the first period, or who had no improvement by week 6, was required to switch to rescue therapy during the crossover period.

### Assessment tools

- **INCAT disability scale**

This disability measure was applied to capture daily activities such as dressing the upper part of the body, doing and undoing buttons and zips, washing and brushing hair, and handling coins. Each item is scored as being 'not affected', 'affected but not prevented', or 'prevented'. The leg scale measures problems with walking, taking into account the use of aids. The INCAT scale ranges from 0 (no signs of disability) to 10 (most severe disability score).<sup>2</sup> For the purposes of the ICE study, the adjusted INCAT disability scale was used, meaning that a change in the upper limb part of this scale from 0 and 1 or vice versa was not considered to be clinically meaningful and was therefore not included in the analyses as being a responder.

- **Vigorimeter**

The Vigorimeter (Martin, Tuttlingen, Germany) quantifies grip strength on a manometer that is connected to a middle-sized rubber bulb through rubber tubing (score range, 0–160 kPa).<sup>4,5</sup> The average value of three assessments at each visit was used for the purposes of the current study.

- **Study procedure**

Both the INCAT disability score and grip strength were assessed at each visit during the first period. Grip strength examination was performed under pre-defined standard conditions.<sup>6</sup> All participating centres were trained prior to the start of the study on how to assess the measures to ensure consistency of application.

### Statistics

- **Early response analyses**

These analyses assessed whether grip strength as an easy and practical method could be used as a diagnostic tool in individual subjects after only one or two courses of treatment by examining statistical differences at the early visits between the two groups (IGIV-C and placebo). Grip strength data were compared with INCAT disability scale results.

- **Changes in global health**

The assessment of global change in health used the answers to question number 2 of the Short Form 36 questionnaire (Medical Outcomes Trust, Boston, MA; SF-36),

which asks patients to compare their global health rating at study entry and at week 24 of the first period.<sup>7</sup> This question is not used in scoring the domains and summary measure of the SF-36. The following question was asked to patients: 'Compared to 1 year ago, how would you rate your health?' Patient response options included five categories: 'much better', 'somewhat better', 'about the same', 'somewhat worse', or 'much worse'.

- **Converting group response to individual responses for grip strength using the minimum clinically important difference concept**

Grip strength measure is a continuous variable, and the statistically significant group change does not necessarily imply a meaningful difference for individual patients. In the ICE study, a  $\geq 1$ -point improvement in the INCAT scale was used as the primary measure of a clinically meaningful change.<sup>2</sup> Grip strength findings can also be converted from a mean group response to an individual response assessment. For the individual response assessment, the grip strength measurement at each visit for each subject was evaluated against a cut-off value to determine if there had been a clinically meaningful response [particularly examining improvement in the IGIV-C group as well as rate of deterioration (the so-called 'meaningful negative responders') in the placebo group]. The cut-off value was based on the minimum clinically important difference (MCID) universal rule of  $0.5 \times \text{SD}$  as was proposed by Sloan and colleagues.<sup>8</sup> In another study, we estimated the SD from the baseline measurement of all CIDP subjects in the ICE study, and MCID was 8 kPa for the Vigorimeter.<sup>9</sup> This MCID value was used to determine the proportion of patients in both treatment groups who were categorized as responders.

- **Correlation and agreement between INCAT disability scale responses and grip strength responses**

The correlation between these two measures was assessed to determine whether grip strength could indirectly provide information on patients' daily functionality, as has been suggested previously.<sup>5</sup> The agreement between the INCAT disability scale responses and the grip strength (dominant hand) responses was also analysed. The kappa coefficient was calculated for assessing the agreement. In addition, the proportion of subjects with clinically meaningful deterioration assessed by INCAT and grip strength was compared and tested using the McNemar's test.

- **Statistical measures of responsiveness**

INCAT disability scale and grip strength were measured at baseline, day 16, week 3, week 6, and every 3 weeks up to week 24. The data up to week 6 were utilized to assess which of these scales was more sensitive in detecting early responses. The standardized response mean (SRM) and area under the curve (AUC) approaches were used to assess responsiveness.<sup>5</sup>



For each scale SRM is calculated as the mean change divided by the standard deviation in change scores. AUC is calculated using the linear trapezoidal rule

$$AUC_{0-week6} = \sum_{i=2}^4 \frac{(SRM_i + SRM_{i-1}) (D_i - D_{i-1})}{2}$$

where  $SRM_i$  is the SRM at visit  $i$ ,  $D_i$  is the number of days at visit  $i$  at baseline, day 16, week 3, and week 6.

- **Statistical description**

Data were evaluated from the intention-to-treat population, defined as all randomized subjects. Data obtained in the first period were used for the purposes of the current paper. A last observation carried forward approach was applied for the change from baseline variables. Analysis of covariance models were used for group response analyses, with treatment and region as fixed effects and baseline measure as covariate. For categorical variables, treatment differences were compared using chi-square tests. Correlation between INCAT disability scale values and grip strength scores was investigated using Pearson's correlation test. For responsiveness assessment, the SRM was calculated as the mean change divided by the SD in change scores and had no unit. Subsequently, AUC were calculated for SRMs using the trapezoidal rule. The kappa statistic was applied to determine the agreement between the proportion of patients showing a response on both measures. All statistical analyses were performed using SAS® version 9.2 (SAS Institute, Cary, NC).

## Results

### General characteristics

One hundred and seventeen individuals were randomized: 59 patients received IGIV-C and 58 patients received placebo. Baseline mean values for grip strength and INCAT disability scale were similar between the two treatment groups.<sup>3</sup>

### Early group response analyses

Statistically significant group response improvement differences were observed at week 6 for both measures (grip strength and INCAT disability scale), in favour of patients treated with IGIV-C. However, only grip strength (dominant hand) measured with the Vigorimeter achieved statistically significant differences at earlier visits (day 16 and week 3) in favour of the IGIV-C group (Table 1). Grip strength (non-dominant hand) showed similar results where statistically significant differences ( $P = 0.043$ ) were observed as early as week 3.

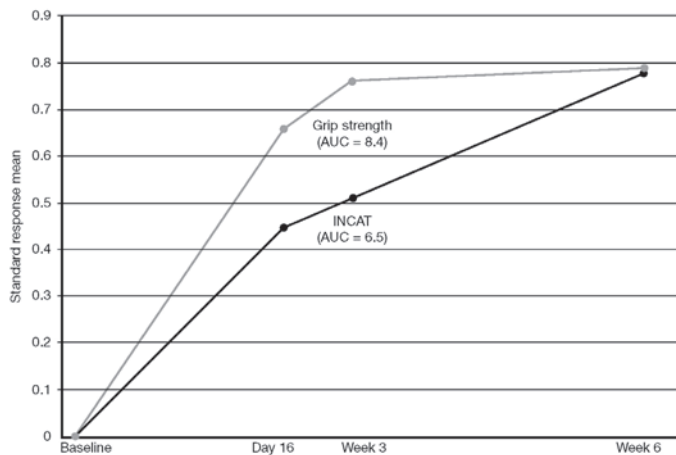
**Table 1.** Change from baseline to day 16, weeks 3, 6, and 24 (endpoint) of first treatment period for adjusted INCAT score and grip strength

Parameter	Visit/change from baseline	IGIV-C		Placebo		Treatment difference (95% CI)	P-value
		N	Mean $\pm$ SD	N	Mean $\pm$ SD		
Adjusted INCAT score	Day 16	57	-0.4 $\pm$ 1.0	57	-0.1 $\pm$ 0.7	-0.31 [0.61, 0.00]	0.051
	Week 3	57	-0.6 $\pm$ 1.1	51	-0.3 $\pm$ 0.8	-0.22 [-0.55, 0.11]	0.196
	Week 6	55	-1.0 $\pm$ 1.3	50	-0.4 $\pm$ 0.9	-0.59 [-1.00, -0.17]	0.006
	Endpoint	59	-1.1 $\pm$ 1.8	58	-0.3 $\pm$ 1.3	-0.72 [-1.27; -0.18]	0.010
Grip strength (dominant hand)	Day 16	56	7.9 $\pm$ 12.1	57	1.5 $\pm$ 14.7	5.98 [1.04, 10.91]	0.018
	Week 3	57	9.0 $\pm$ 11.8	51	2.0 $\pm$ 13.8	5.72 [0.88, 10.56]	0.021
	Week 6	54	11.9 $\pm$ 15.1	50	3.3 $\pm$ 12.7	7.53 [2.21, 12.85]	0.006
	Endpoint	57	13.2 $\pm$ 19.3	58	1.5 $\pm$ 15.6	10.94 [4.65, 17.22]	< 0.001

### Changes in global health

Figure 1 shows the relationship between change in grip strength and the five categories of the global change in health as assessed using question number 2 of the SF-36. The figure shows that subjects with more improvement in grip strength tended to rate their improvement in global health better than those with less improvement in grip strength (Spearman's correlation coefficient:  $r = 0.47$ ,  $P < 0.001$ ).

**Figure 1.** Change in grip strength (dominant hands) related to global health changes



**Legend to Figure 1:** Changes in global health were assessed using question number 2 of the SF-36 questionnaire: 'compared to 24 weeks ago, how would you rate your health?'

### Converting group response to individual response for grip strength using MCID

The cut-off value for grip strength was based on the MCID concept of  $0.5 \times \text{SD}$  as the universal rule and was 8 kPa.<sup>8,9</sup> Table 2 shows the proportion of individual meaningful

improvement responses using the MCID cut-off. There were more responders in IGIV-C group than the placebo group at each of the first three visits (day 16, week 3, and week 6). The results are similar to the proportion of individual responses using the adjusted INCAT score (data not shown).

**Table 2.** Individual clinically meaningful improvement response in grip strength (dominant hand) by treatment arm using the universal rule for the MCID (0.5 ´ SD)

Visit	IGIV-C		Placebo		P-value
	N	Response (%)	N	Response (%)	
Day 16	56	37.5	57	21.1	0.055
Week 3	56	39.3	51	23.5	0.080
Week 6	54	55.6	50	24.0	0.001
Last measurement in first period	57	50.9	58	25.9	0.006

**Legend to Table 2:** IGIV-C, immune globulin; MCID, minimum clinically important difference. Response is defined as change in grip strength of at least 8 kPa.

In the placebo group, a higher proportion of patients with a meaningful deterioration was captured by grip strength (negative responders showing > 8 kPa reduction) within the first 6 weeks of the study when compared with the INCAT score findings (Table 3). At day 16, 13 patients (22.8%) were negative responders with the grip strength and only 4 (7%) with the INCAT disability score (P = 0.003); at week 3: grip strength; INCAT = 8 (15.7%) versus 1 (2%) (P = 0.008), at week 6: grip strength; INCAT = 10 (20%) versus 4 (8%) (P = 0.058) (Table 3).

**Table 3.** Clinically meaningful deterioration assessed by the INCAT disability score and grip strength in the placebo group

Visit	INCAT score significant negative response	Grip strength significant negative response N (%)		P-value
		No	Yes	
Day 16 (N = 57)	No	44 (77.2)	9 (15.8)	0.003
	Yes	0	4 (7.0)	
Week 3 (N = 51)	No	43 (84.3)	7 (13.7)	0.008
	Yes	0	1 (2.0)	
Week 6 (N = 50)	No	38 (76.0)	8 (16.0)	0.058
	Yes	2 (4.0)	2 (4.0)	

**Legend to Table 3.** ‘Significant negative responder’ was defined as a patient showing deterioration defined as (i) 1-point increase (= worsening) on the INCAT; (ii) 8-point deterioration on the Vigorimeter.

**Correlation and agreement between INCAT disability scale responses and grip strength responses**

The correlation between the INCAT disability scale values and grip strength scores was highly significant (at baseline: r = 0.44; at first period endpoint: r = 0.61; P < 0.001). The results in tables 3 and 4 show a fair agreement between INCAT disability scale responses and the grip strength score responses, with agreements ranging from

63.6% to 84.3%. However, there were some discordance, where INCAT changed by  $\geq 1$  point but grip strength improved by  $< 8$  kPa, or grip strength improved by  $> 8$  kPa but there was no improvement in INCAT disability score.

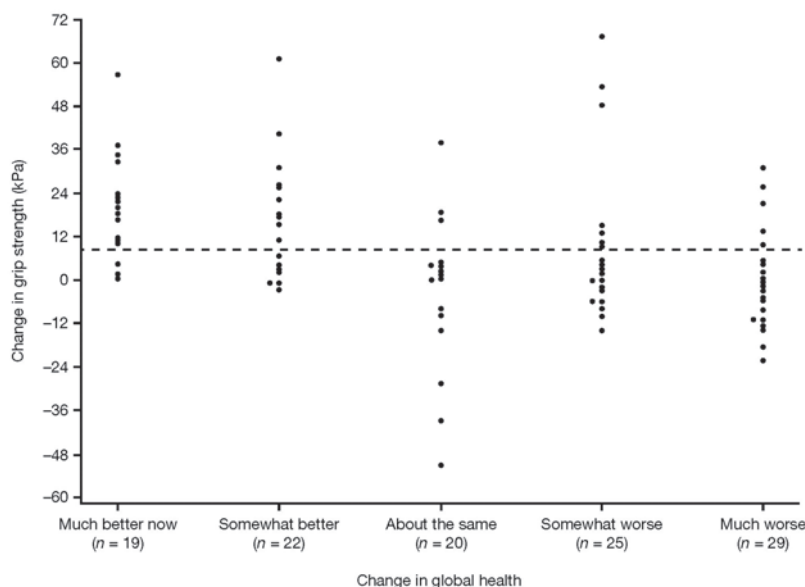
**Table 4.** Agreement between the INCAT disability score and grip strength responses

Visit	INCAT score response	Grip strength response N (%)		Kappa statistic
		No	Yes	
Day 16 (N = 113)	No	72 (63.7)	19 (16.8)	0.359
	Yes	8 (7.1)	14 (12.4)	
Week 3 (N = 107)	No	57 (53.3)	23 (21.5)	0.110
	Yes	16 (15.0)	11 (10.3)	
Week 6 (N = 104)	No	42 (40.4)	12 (11.5)	0.380
	Yes	20 (19.2)	30 (28.8)	

### Statistical measure of responsiveness

Among the IGIV-C group, larger SRMs were observed at day 16 and week 3 for grip strength score than for INCAT disability score (Figure 2). By week 6, both scales had similar SRMs (0.78 for INCAT disability score and 0.79 for grip strength). The AUC over the 6-week period was 8.4 for grip strength and 6.5 for INCAT disability score.

**Figure 2.** Standardized response mean (SRM) versus treatment duration for inflammatory neuropathycouse and treatment (INCAT) score and grip strength.



**Legend to Figure 2:** SRM values are shown, together with corresponding area under the curve, for INCAT score and grip strength for subjects treated with IGIV-C as part of the early response comparison

## Discussion

This study provides data showing the clinical value of using the Martin Vigorimeter, an easily applicable tool for the assessment of grip strength in patients with CIDP. Statistically significant, clinically meaningful improvement in CIDP patients being treated with IGIV-C, compared with those given placebo, was captured with the Vigorimeter, and these changes were seen as early as day 16 of follow-up in the ICE study. By contrast, statistically significant changes were not seen with the INCAT disability score until week 6 (Table 1).<sup>3</sup> Relevant changes in grip strength showed a good correlation with the patients' own judgement of change in their global health measured using question number 2 of the SF-36 (Figure 1).<sup>7</sup> Generally, individuals with an improvement in grip strength reported better improvement in global health and vice versa. The latter findings are considered a heuristic form of responsiveness.<sup>10</sup> In addition, the responsiveness assessed by the non-unit SRM and AUC indicated that grip strength is a more sensitive measure of response to treatment at early visits than the INCAT disability score. An extension of the responsiveness comparison between the two outcome measures beyond week 6 was limited because of the conditional crossover design of the study.<sup>2</sup>

The concept of MCID has been thoroughly addressed in a recent paper using the ICE trial data.<sup>9</sup> For both the INCAT disability scale and grip strength scores, higher proportions of responders were seen in the IGIV-C group than the placebo group for the anchor-based and distribution-based MCID used. In the current study, a higher proportion of patients reached the cut-off MCID value of 8 kPa in the IGIV-C group, thus showing clinically meaningful improvement (Table 2). Hence, in addition to statistically significant comparison findings for the groups, these findings demonstrate higher response rates for both measures in favour of the IGIV-C group at the individual level.

The Martin Vigorimeter also demonstrated higher sensitivity at the other end of the spectrum of changes: this tool significantly captured as early as day 16 more patients in the placebo group showing a meaningful deterioration (MCID cut-off > 8 kPa) when compared with the INCAT findings (Table 3). To our knowledge, these findings have not been published previously and could be used as an early indication for additional therapy if patients continue to deteriorate despite having received the first full IVIG dosage. However, future studies should however focus on the impact of a second IVIG dosage in these patients and determine whether they respond more rapidly. In a recent paper it was demonstrated that some subjects respond (as determined by the INCAT score) to IVIG therapy after one infusion course while other patients required two infusion courses to see a response.<sup>11</sup> This also holds true with the grip strength data, which showed the proportion of responses significantly increased from 37.5% at day 16 and 39.3% at week 3 to 55.6% at week 6 (Table 2). However, further studies are needed to explore the factors that might be predictive of needing more IVIG

courses before a response to therapy is observed. These observations could also be of particular interest for the design of future maintenance trials in patients with CIDP.

All these findings indicate that, in addition to being simple, quick, and easy to use and fulfilling the clinimetric requirements of being valid, reliable, and responsive, the Vigorimeter detects responses quickly and is related to patients' own judgement of their clinical state (Figure 1).<sup>5</sup> Grip strength also demonstrated an acceptable correlation with the INCAT disability scale with moderate response agreement (Tables 3 and 4), which conforms previous reports.<sup>5,12,13</sup> However, there were also discordances between the two measures where INCAT score showed improvement without grip strength improvement and vice versa (Tables 3 and 4). This may be explained by some subjects having greater lower extremity disability changes that INCAT can detect, but grip strength cannot. Also, changes in grip strength (located at the impairment level of assessing outcome) do not always translate to a change in disability. Hence, the moderate agreement indicates that other factors exist that contribute to disability.

There are some methodological issues that should be addressed. Although the Vigorimeter is easily applicable, fulfills all clinimetric requirements and correlating moderately with disability, it only assesses impairment distally and does not provide information on clinically important proximal muscle groups. Dynamometric devices have been suggested to overcome this deficit, making it possible to quantify possible proximal weakness.<sup>14,15</sup> Although seemingly more comprehensive in their applicability these devices have a higher burden to patients and may even cause some discomfort.<sup>5</sup> In addition, as has been demonstrated for the Vigorimeter, no other grip strength device has been subjected to a thorough examination to determine its scientific soundness in these illnesses. Although the use of the Vigorimeter in daily practice is suggested, physicians should also focus on other aspects contributing to problems in daily living and social functioning in patients with CIDP.<sup>16</sup>

Despite these observations, grip strength, reflecting distal strength and upper limb function, has shown its usefulness as a prognostic indicator of clinical and functional recovery and is useful in monitoring the effect of treatment.<sup>12,13,17</sup>

In conclusion, grip strength assessment with the Vigorimeter is a suitable outcome measure that is recommended for routine follow-up studies of patients with CIDP.

## References

- 1 Merckies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 2 Hughes, R., Bensa, S., Willison, H. *et al.* Randomized controlled trial of intravenous immunoglobulin versus oral prednisolone in chronic inflammatory demyelinating polyradiculoneuropathy. *Ann Neurol* **50**, 195-201 (2001).
- 3 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet Neurol* **7**, 136-144 (2008).
- 4 Fünfgeld, E. W. Vigorimeter: Zur Kraftmessung der Hand und zur Simulationsprüfung. *Dtsch Med Wochenschr* **49**, 2214-2216 (1966).
- 5 Merckies, I. S., Schmitz, P. I., Samijn, J. P. *et al.* Assessing grip strength in healthy individuals and patients with immune-mediated polyneuropathies. *Muscle Nerve* **23**, 1393-1401 (2000).
- 6 Therapists, A. S. o. H. *Clinical Assessment Recommendations*. (American Society of Hand Therapists, 1992).
- 7 Ware, J. E. J., Kosinski, M., and Gandek, B. *SF-36 Health Survey: Manual and Interpretation Guide*. (Lincoln, RI: QualityMetric Incorporated, 2000).
- 8 Sloan, J., Symonds, T., Vargas-chanes, D. *et al.* Practical Guidelines for Assessing the Clinical Significance of Health-Related Quality of Life Changes within Clinical Trials. *Drug Information Journal* **37**, 23-31 (2003).
- 9 Merckies, I. S., van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *J Neurol Neurosurg Psychiatry* **81**, 1194-1199 (2010).
- 10 Liang, M. H. Evaluating measurement responsiveness. *J Rheumatol* **22**, 1191-1192 (1995).
- 11 Latov, N., Deng, C., Dalakas, M. C. *et al.* Timing and course of clinical response to intravenous immunoglobulin in chronic inflammatory demyelinating polyradiculoneuropathy. *Arch Neurol* **67**, 802-807 (2010).
- 12 Orndahl, G., Sellden, U., Hallin, S. *et al.* Myotonic dystrophy treated with selenium and vitamin E. *Acta Med Scand* **219**, 407-414 (1986).
- 13 Rhind, V. M., Bird, H. A., and Wright, V. A comparison of clinical assessments of disease activity in rheumatoid arthritis. *Ann Rheum Dis* **39**, 135-137 (1980).
- 14 Dyck, P. J. Dynametric assessment in CIDP. *Muscle Nerve* **39**, 421-422 (2009).
- 15 Harbo, T., Andersen, H., and Jakobsen, J. Acute motor response following a single IVIG treatment course in chronic inflammatory demyelinating polyneuropathy. *Muscle Nerve* **39**, 439-447 (2009).
- 16 Merckies, I. S., Hughes, R. A., Donofrio, P. *et al.* Understanding the consequences of chronic inflammatory demyelinating polyradiculoneuropathy from impairments to activity and participation restrictions and reduced quality of life: the ICE study. *J Peripher Nerv Syst* **15**, 208-215 (2010).
- 17 Sunderland, A., Tinson, D., Bradley, L. *et al.* Arm function after stroke. An evaluation of grip strength as a measure of recovery and a prognostic indicator. *J Neurol Neurosurg Psychiatry* **52**, 1267-1272 (1989).

# Modifying the Medical Research Council grading system through Rasch analyses

*Els Karla Vanhoutte<sup>1</sup>, MD; Catharina Gerritina Faber<sup>1</sup>, MD, PhD; Sonja Ingrid van Nes<sup>2</sup>, MD; Bart Casper Jacobs<sup>3</sup>, MD, PhD; Pieter Antoon van Doorn<sup>2</sup>, MD, PhD; Rinske van Koningsveld<sup>3</sup>, MD, PhD; David Reid Cornblath<sup>4</sup>, MD, PhD; Anneke Jelly van der Kooij<sup>5</sup>, MD, PhD; Elisabeth Aviva Cats<sup>6</sup>, MD; Leonard Hendrik van den Berg<sup>6</sup>, MD, PhD; Nicolette Claudia Notermans<sup>6</sup>, MD, PhD; Willem Lodewijk van der Pol<sup>6</sup>, MD, PhD; Mieke Catharina Elisabeth Hermans<sup>1</sup>, MD; Nadine Anna Maria Elisabeth van der Beek<sup>3</sup>, MD; Kenneth Craig Gorson<sup>7</sup>, MD, PhD; Marijke Eurelings<sup>8</sup>, MD, PhD; Jeroen Engelsman<sup>8</sup>, MD; Hendrik Boot<sup>8</sup>, MD, PhD; Ronaldus Jacobus Meijer<sup>8</sup>, MD; Giuseppe Lauria<sup>9</sup>, MD, PhD; Alan Tennant<sup>10</sup>, PhD; Ingemar Sergio José Merkies<sup>1,8\*</sup>, MD, PhD; on behalf of the PeriNomS study group*

<sup>1</sup>Department of Neurology, Maastricht University Medical Centre, School of Mental Health and Neuroscience, Maastricht, the Netherlands; <sup>2</sup>Department of Neurology, Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands; <sup>3</sup>Department of Neurology, Elkerliek Hospital Helmond, the Netherlands; <sup>4</sup>Department of Neurology, Johns Hopkins Hospital and School of Medicine, Baltimore, USA; <sup>5</sup>Department of Neurology, Academic Medical Centre, Amsterdam, the Netherlands; <sup>6</sup>the Rudolf Magnus institute of neuroscience, University Medical Centre Utrecht, Utrecht, the Netherlands; <sup>7</sup>Department of Neurology, Caritas St. Elizabeth's Medical Centre, Tufts University School of Medicine, Boston Massachusetts, USA; <sup>8</sup>Department of Neurology, Spaarne Hospital, Hoofddorp, the Netherlands; <sup>9</sup>Neuromuscular Diseases Unit, IRCCS Foundation, "Carlo Besta", Milan, Italy; <sup>10</sup> Department of Rehabilitation Medicine, Faculty of medicine and health, University of Leeds, Leeds, UK.



## Abstract

The Medical Research Council grading system has served through decades the medical field for the evaluation of muscle strength and has been recognized as a cardinal feature of daily neurological, rehabilitation, and general medicine examination of patients, despite being respectfully criticized due to the unequal width of its response options. No study has systematically examined through modern psychometric approach whether physicians are able to properly use the Medical Research Council grades.

The objectives of the study were (i) to investigate physicians' ability to discriminate among the Medical Research Council categories in patients with different neuromuscular disorders and with various degrees of weakness through thresholds examination using Rasch analysis as a modern psychometric method; (ii) to examine possible factors influencing physicians' ability to apply the Medical Research Council categories through differential item function analyses; and (iii) to examine whether the widely used Medical Research Council 12 muscles sumscore in patients with Guillain-Barré syndrome and chronic inflammatory demyelinating polyradiculoneuropathy would meet Rasch model's expectations. A total of 1065 patients were included from nine cohorts with the following diseases: Guillain-Barré syndrome (n=480); myotonic dystrophy type-1 (n=169); chronic inflammatory demyelinating polyradiculoneuropathy (n=139); limb-girdle muscular dystrophy (n=105); multifocal motor neuropathy (n=102); Pompe's disease (n=62); monoclonal gammopathy of undetermined related polyneuropathy (n=8). Medical Research Council data of 72 muscles were collected. Rasch analyses were performed on Medical Research Council data for each cohort separately and after pooling data at the muscle level to increase category frequencies, and on the Medical Research Council sumscore in patients with Guillain-Barré syndrome and chronic inflammatory demyelinating polyradiculoneuropathy. Disordered thresholds were demonstrated in 74-79% of the muscles examined, indicating physicians' inability to discriminate between most Medical Research Council categories. Factors such as physicians' experience or illness type did not influence these findings. Thresholds were restored after rescoring the Medical Research Council grades from six to four options (0: paralysis, 1: severe weakness, 2: slight weakness, 3: normal strength). The Medical Research Council sumscore acceptably fulfilled Rasch model expectations after rescoring the response options and creating subsets to resolve local dependency and item bias on diagnosis.

In conclusion, a modified, Rasch-built four response category Medical Research Council grading system is proposed, resolving clinicians' inability to differentiate among its original response categories and improving clinical applicability. A modified Medical Research Council sumscore at the interval level is presented and is recommended for future studies in Guillain-Barré syndrome and chronic inflammatory demyelinating polyradiculoneuropathy.

## Introduction

In 2005, a historical essay tracing the history of scoring and summation of neuromuscular weakness as part of daily neurological practice.<sup>1</sup> Mitchell and Lewis initiated the practice of alphanumerical scoring of neurological signs in the 19th century.<sup>2</sup> However, it was Lovett, an orthopaedic surgeon who introduced an ordinal scoring of muscle weakness that formed the basis for the Mayo Clinics and Medical Research Council (MRC) manual muscle testing grading systems, of which the MRC system is most widely used<sup>1,3</sup>. Its worldwide recognition is most probably due to its simplicity and drawings illustrating how limb muscles should be tested. Through the decades, various versions have been published that aimed to improve the methods for muscle examination. The 2010 edition of *Aids to the Investigation of Peripheral Nerve Injuries. Medical Research Council: Nerve Injuries Research Committee* was recently presented on behalf of the guarantors of Brain embracing a historical review and appreciation for its nurtures through the decades.<sup>4</sup> Despite being the most cardinal feature of daily neurological practice, the MRC scale has been respectfully criticized due to the unequal width of its categories, with Grades 1, 2, and 3 being too narrow, and 4 being too broad, often leading to attempts to modify the scale.<sup>1,5-8</sup>

One of the most common sources of improper use of any outcome measure concerns the inconsistent use of the response options that correspond to the scale's items.<sup>9</sup> This results in what is known as "reversed or disordered thresholds". The term threshold refers to the point between two adjacent response categories where either response is equally probable. In the case of the MRC scale, a threshold would be the point between two adjacent categories, such as between MRC Grades 2 and 3. Disordered thresholds occur when physicians have difficulty consistently discriminating among the MRC grades in patients with various degrees of muscle weakness. Surprisingly, no study has systematically examined the appropriateness of the MRC scale using modern psychometric techniques.

The objectives of this study were: (i) to examine the applicability and discriminative capacity of physicians using the MRC grades in patients with various neuromuscular illnesses with different degrees of muscle weakness. We questioned whether physicians could demonstrate a fairly uniform MRC grades' ordered thresholds pattern along the Rasch scale continuum, since previous reports suggested human's inability to differentiate between more than four response options;<sup>10,11</sup> (ii) to investigate the influence of factors possibly affecting the proper use of the MRC grades in clinical practice (such as physician's clinical experience). For these two objectives, the Rasch method as a modern psychometric vehicle was used, solely focusing on threshold and item bias examinations;<sup>9,12</sup> and (iii) since Guillain-Barré syndrome and chronic inflammatory demyelinating polyradiculoneuropathy (CIDP) are potentially treatable illnesses and the MRC sum score has been often used as an outcome measure to determine efficacy in these illnesses, we have chosen to examine whether this multi-

item scale would fulfil all Rasch's model expectations in patients with Guillain-Barré syndrome and CIDP and if not, to propose changes to improve its use.<sup>13-16</sup>

Patients and methods

Patients

The MRC grades of various muscles were collected from different neuromuscular seminal studies published in the last two decades. Most of these studies have guided the worldwide neurological community in understanding the clinical and therapeutic pattern of these illnesses. A total of 1065 patients (GBS: n=480; myotonic dystrophy type 1: n=169; CIDP: n=139; limb-girdle muscular dystrophy: n=105; multifocal motor neuropathy: n=102; Pompe's disease: n=62; monoclonal gammopathy related polyneuropathy of undetermined significance: n=8) were included (Table 1 and Supplementary Table 1).<sup>14-22</sup> The initial MRC data of all patients from the above-mentioned cohorts were selected for the purposes of the current study. All patients met their international criteria for their illness.<sup>23-28</sup> The diagnosis 'monoclonal gammopathy related polyneuropathy of undetermined significance' was established after excluding all possible causes for the gammopathy and polyneuropathy.<sup>29</sup> For all studies, consent was obtained according to the Declaration of Helsinki and approval was obtained by the Ethical Committee of the institution in which the original study was performed.

Table 1. Basic characteristics of patients with neuromuscular disorders

Study/disorder	Patients examined (n)	Age mean years (SD), range	Gender		symptoms duration mean years (SD), range
			Female (%)	Male (%)	
INCAT study	113	54.3 (15.1), 14-84	54 (47.8)	59 (52.2)	6.9 (3.1), 0.5-28
Dutch GBS trial 1992	147	47.5 (19.2), 5-81	71 (48.3)	76 (51.7)	-
Dutch GBS trial 2004 + GBS pilot study 1994	250	50.5 (20.1), 7-89	109 (43.6)	141 (56.4)	-
DM1 patients	169	43.5 (11.5), 18-69	83 (49.1)	86 (50.9)	5.3 (6.9), 0-34
MMN patients	102	54.3 (12.1), 26-79	76 (74.5)	26 (25.5)	11.8 (8.2), 0.2-43
Pompe's patients	62	48.1 (11.9), 25.1-71.7	29 (46.8)	33 (53.2)	7.9 (9.3), 0-30.5
LGMD patients	105	37.8 (15.6), 3-70	64 (61.0)	41 (39.0)	21.0 (14.5), 0-58
ICE CIDP study	117	51.6 (16.5), 18-83	40 (34.2)	77 (65.8)	5.3 (6.2), 0.2-34.3

**Legend to Table 1:** In the INCAT studies, a total of 113 patients were examined (Guillain-Barré syndrome, n = 83; CIDP, n = 22, and monoclonal gammopathy related polyneuropathy of undetermined significance, n = 8).

Assessment scale

The MRC grading system provides the following grades: 0, paralysis; 1, only a trace or flicker of muscle contraction is seen or felt; 2, muscle movement is possible with gravity eliminated; 3, muscle movement is possible against gravity; 4, muscle strength

is reduced, but movement against resistance is possible and 5, normal strength. The MRC grades of the following six muscle pairs comprise the MRC sum score for Buillain-Barré syndrome and CIDP: upper arm abductors, elbow flexors, wrist extensors, hip flexors, knee extensors, and foot dorsal flexors.<sup>13</sup> In the remaining cohorts (monoclonal gammopathy of undetermined significance related polyneuropathy, multifocal motor neuropathy, myotonic dystrophy type-1, Pompe's disease, limb-girdle muscular dystrophy), the muscles groups evaluated represented the clinical picture of each illness (see Supplemental Table 1 for available muscles per cohort).

## Rasch analysis

### Rationale for using Rasch method

In health care, outcome measures often consist of ordinal multi-item questionnaires, based on the classical test theory.<sup>30</sup> Concerns have been raised about inappropriate analysis of the generated summed scores that are erroneously assumed to be at the interval level.<sup>30-32</sup> The ability of a scale to provide fundamental measurements should be established before the more commonly reported psychometric attributes such as being simple, valid, reliable, and responsive.<sup>9,33</sup> Modern scientific methods have been adopted to overcome the shortcomings of traditional measurements. One of the most widely used modern techniques is the Rasch method that transforms ordinal obtained scores into interval-level variables, and whose fit of data satisfies numerous checkpoints as part of model expectations.<sup>9,12,33</sup>

In the current study setting, the Rasch model assumes that a patient with less weakness (thus more strength) will have a greater chance of receiving a higher MRC grade by the examining physician. A comprehensive description of the Rasch analysis specifically for neurologists is provided elsewhere.<sup>9,34,35</sup> Briefly, the Rasch model shows what should be expected in response to ordinal items if interval scaling is to be achieved. For this, the following criteria should be fulfilled:

- Thresholds examination: when using items with more than two response categories, as for the MRC grades, proper ordering of the response options should be verified using category probability curves for each muscle group examined, since this will reflect the ability of physicians to use the MRC in a proper way.<sup>36</sup> Ordered thresholds are where the transition (threshold) between categories map on to the underlying construct in the expected manner. Thus the transition between categories, (e.g.1-2 and 2-3) reflects increasing levels of muscle strength (see Figure 1, top). Disordered thresholds can occur when physicians use the response options inconsistently, and this inconsistency can be a source of misfit to model expectations. The difficulty discriminating between response options may be a result of too many options, or where the labelling of the options is confusing, both of which may lead to misinterpretation.

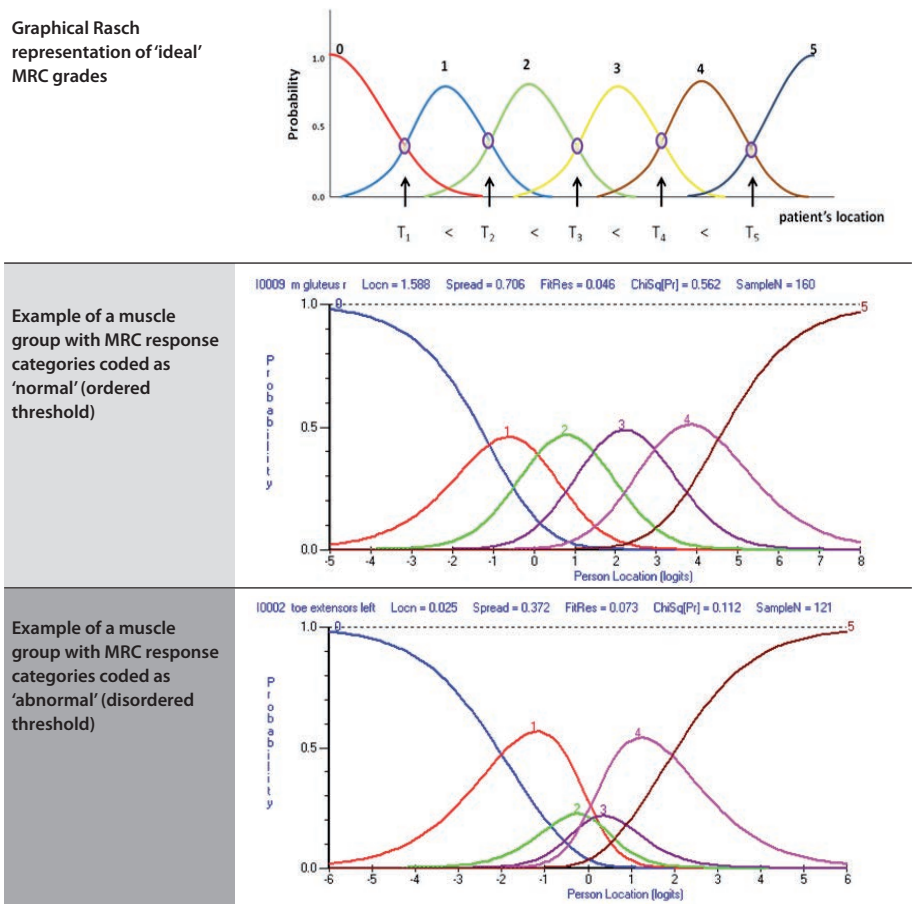
- Fit statistics: fit statistics give an indication of how well the items fit the expected ordering required by the model. This ordering is a probabilistic version of Guttman Scaling.<sup>37</sup> There are two general categories for detecting misfit: overall (summary) misfit, using the entire response matrix, and the individual fit (examining all items and all persons individually). At the summary level the overall mean residual values for both persons and items can be calculated. These values are expressed as a z-score with a mean of 0 and a SD of 1, values of which indicate perfect fit to model expectations.<sup>34,38</sup> The summary item-trait interaction statistic reflects the fit of the observed data to the model's expectations and is represented by the chi-square. This statistic gives an indication of the invariance of the ordering of items across patients with different levels of muscle strength. A significant chi-square indicates a failure to retain this ordering. Besides the overall fit residuals, individual item-chi square and item and person residuals can be calculated.<sup>34,38,39</sup>
- Item bias: response to an item should not vary between groups (e.g.: males versus females), given the same level of the underlying trait (e.g. muscle strength). We assessed item bias (Differential Item Functioning) on the MRC data for various available person factors. A panel (I.S.J.M. and C.G.F.) have studied the range of the factors age, disease duration, and physician's experience in the available cohorts. Subsequently, these factors were categorized into subgroups for item bias analyses, aiming for an equivalent distribution of participants among the subgroups (25-33% per subgroup).
- Local dependency: local dependency arises when items are linked such that the response on one item is dependent upon the response to another. Item sets with correlations  $> 0.3$  are considered a source of misfit to the model.<sup>9</sup>
- Unidimensionality: the Rasch model assumes unidimensionality and consequently post-hoc tests are included in the analysis to ensure that this assumption holds. These tests involve a comparison of person estimates (of muscle strength) based upon two set of items identified from the first principal component analysis of the residuals. The estimates for every individual are compared by a *t*-test, and where  $< 5\%$  of these comparisons are significantly different, this is taken to support the assumption of unidimensionality.<sup>40</sup>

### Test procedure

Figure 2 presents a systematic ordering of the analyses performed in the current study. In analyses 1 and 2 (MRC Rasch analyses for each cohort separately and MRC Rasch analyses after pooling data) the following was examined:

- Step 1: the presence of ordered thresholds, thus determining whether the MRC grades for each muscle were ordered reflecting physicians' ability to use these grades properly;
- Step 2: in case of disordered thresholds: to seek for the most optimal modified MRC rescored categories that could serve as a unified tool in manual muscle scoring for

**Figure 1.** MRC response categories related thresholds explained and coded as ‘normal (green box)’ or ‘abnormal (red box)’



**Legend to Figure 1:** MRC response categories related thresholds explained and coded as ‘normal’ (green) or ‘abnormal’ (red). The first row shows the ideal graph representation for proper thresholds for the MRC grades. The first threshold at the intersection between MRC response options 0 and 1 corresponds to 50% chance of choosing between these two adjacent categories. The thresholds should be ordered to obtain an ideal graph: Threshold 1 < Threshold 2 < Threshold 3 < Threshold 4 < Threshold 5. The second and third row give graphical examples of proper threshold ordering (coded as a green box) and disordered threshold (coded as a red box), respectively. T1-T5 = Thresholds 1-5, respectively.

all muscle groups. In order to rescore the MRC categories, the frequency distribution among the categories and the category probability curves were taken into account. The frequency distribution among the categories and the category probability curves were taken into account;

- Step 3: the presence of possible item bias were examined to determine whether factors like physician's experience in the neuromuscular field (would a more experienced physician apply the MRC grades more appropriately than a less experienced physician?) or possible differences between community and university based neurologists might influence the applicability of the MRC grades.

Therefore, in Analyses 1 and 2 the Rasch method was applied only to examine the ability of physicians to use the MRC grading system in a proper way and to determine whether there are factors influencing its use. These analyses were not intended to create a formal Rasch-built MRC sum score for each cohort individually, since some of the cohort samples were relatively small, hence not fulfilling the basic requirements for proper Rasch modeling.<sup>41</sup>

For analyses 2, MRC data were pooled at the muscle level from the various available cohorts and re-subjected to Rasch analysis, thereby controlling for diagnosis as a possible confounder and strengthening the category frequencies for the various muscles.<sup>42</sup>

In analysis 3 (MRC sums core Rasch analysis in Guillain-Barré syndrome/CIDP), the MRC 12 muscles sum score was analysed to determine whether Rasch model expectations would be met. The first two steps for analyses 1 and 2 (see above) were also performed here. Subsequently, since there is no consensus regarding a fixed sequence of steps that must be followed when doing Rasch analyses, our rationale for the following steps were constantly driven by the biggest abnormality seen when studying all subjected data to Rasch, hereby focusing on all aspects that did not meet model expectations (misfit statistics, fit residuals disturbances, under-/overfitting, local dependency > 0.3, and item bias). All steps needed were taken to create a unidimensional scale at the interval level.

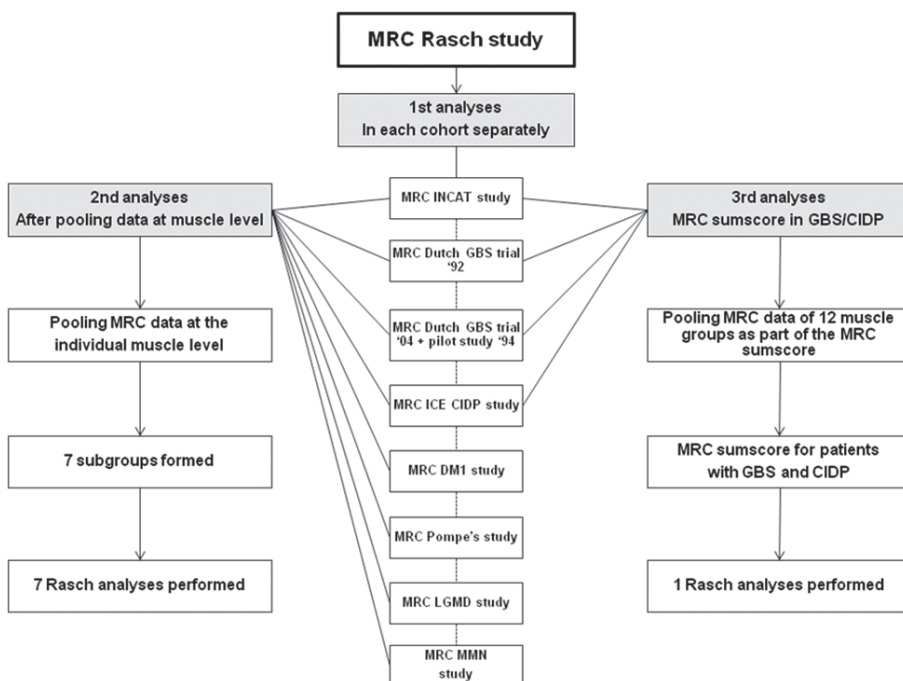
### **Rasch general aspects, person factors, and statistics**

The MRC data of each muscle group was treated as if it was an 'item' that needed to be completed by the patients with response options from 0 to 5 (in the current study setting: a physician completed the 'item') using the Rasch Unidimensional Measurement Model 2020 software.<sup>43</sup>

In Analysis 1 (MRC Rasch in each cohort separately), the following person factors were taken into account (Supplementary Table 2):

- Age: 1, <40 years; 2, 40 to 59 years and 3, ≥60 years;
- gender: 0, female; 1, male;
- type of disease: (a) inflammatory-neuropathy-cause-and-treatment cohort: 1, Guillain-Barré syndrome; 2, CIDP; 3, gammopathy related polyneuropathy; (b) myotonic dystrophy cohort: 1, mild; 2, adult; 3, child/congenital type; and (c) limb-girdle dystrophy cohort: 1, sarcoglycanopathy; 2, calpainopathy; 3, limb-girdle type 1B, 2B and 2I; 4, unclassified;

**Figure 2.** Study algorithm showing a systematic ordering of the analyses performed in the current study



**Legend to Figure 2:** Study algorithm showing a systematic ordering of the analyses performed in the current study. First analyses (Analysis 1): initial MRC Rasch analysis for each individual cohort separately (thus performing a total of 8 individual model analyses); Second analyses (Analyses 2): MRC Rasch analyses after pooling data at the muscle level from available cohorts. Third analyses (Analysis 3): MRC sumscore Rasch analysis in patients with GBS and CIDP. DM1 = myotonic dystrophy type-1; ICE= immune globulin intravenous for chronic inflammatory demyelinating polyneuropathy; LGMD = limb-girdle muscular dystrophy; MMN = multifocal motor neuropathy

- duration of disease: (a) for all cohorts except limb-girdle patients: 1, <5 years; 2, 5-9 years; 3, 10-19 years; 4, ≥20 years; and (b) for limb-girdle cohort: 1, <10 years; 2, 10-19 years; 3, 20-29 years; 4, ≥30 years;
- physician's experience in the neuromuscular field: for the inflammatory-neuropathy cause and treatment studies: 1, <3 years experience; 2, 3-5 years experience; 3, ≥6 years experience; the latter group constituting of senior neuromuscular experts;
- institution; for the Guillain-Barré syndrome trials: 0, community based; 1, university based hospital; and
- country; for the Guillain-Barré syndrome cohort 2004: 1, The Netherlands; 2, Germany; 3, Belgium.

For analyses 2 and 3 (MRC Rasch after pooling data and MRC sum score in Guillain-Barré syndrom/CIDP), the factors studied included (i) age category: 1, <40 years; 2, 40-



59 years; 3,  $\geq 60$  years; (ii) gender: 0, female; 1, male; and (iii) type of disease: depending on the amount of illnesses being pooled together, each illness received a separate code.

For the MRC sum score analysis, the person separation index was also determined, which should be  $\geq 0.7$  for proper group comparison, and a minimum of 0.9 for clinical use.<sup>44</sup> The unrestricted partial credit Rasch model was used. Further analyses were undertaken using Stata 11.0 statistical software for Windows XP.

## Results

### General aspects

A total of 1065 patients with various neuromuscular disorders were included from nine studies. Table 1 presents the patients' characteristics. MRC data on 72 muscle groups were available (Supplementary Table 1: muscle groups assessed per cohort).

### Analyses 1: Initial MRC Rasch analyses for each cohort separately (see Figure 2)

- **Step 1 Thresholds examination**

The obtained data (ordered thresholds coded "green"; disordered coded "red"; see Figure 1 explaining these codes) for each muscle group in each cohort were summed, hereby creating a total of 210 muscle groups examined. A total of 165 (78.6%) muscle groups had disordered thresholds versus 45 (21.4%) with ordered thresholds. The disordered thresholds were particularly seen in the mid-response MRC category area (options 2 to 4).

- **Step 2 Rescoring MRC categories**

A panel of neuromuscular and Rasch researchers studied the category probability curves and category frequencies of the MRC data for each muscle group. Subsequently, all muscle groups were systematically rescored in order to obtain the maximum uniform amount of response options, which turned out to be four categories (instead of six). Of the 210 muscle groups rescored, 182 (86.7%) had ordered thresholds and 28 (13.3%) were still disordered. Sixteen of these disordered muscle groups were distally located (finger spreaders, flexors, and extensors, grip strength, wrist extensors and flexors, foot dorsal and plantar flexors). All disordered muscle groups except two were found in the two cohorts with the lowest amount of patient's records (multifocal motor neuropathy:  $n = 102$  and Pompe's disease,  $n = 62$ ).

- **Step 3 Item bias examination**

Eight selected person factors were used to examine possible item bias on the available muscle groups (see Supplemental Table 2 for available factors per cohort). Before rescoring, a total of 806 muscle groups (96.9%) were free of item bias, thus not being influenced by person factors like physicians' experience. Item bias was only found in 26 muscles (3.1%; on person factor gender: 11 muscle groups had

uniform differential item functioning, on disease type, eight had uniform, on disease duration: tow uniform and one non-uniform, on physician's experience: two uniform, on country: one uniform, and on age: one muscle group had uniform differential item functioning). Differential item functioning findings did not change after rescoring at the individual cohort level.

## Analysis 2: MRC Rasch analyses after pooling data

### • Step 1 Thresholds examination:

Similar findings were seen in the pooled data analyses. Of the 72 muscles examined, a total of 53 muscle groups (73.6%) had disordered threshold, particularly in the mid-categories (Table 2, 'before rescoring').

**Table 2.** Results before and after rescoring the response options from 6 to 4 categories with corresponding threshold locations

	Studies (n)	Pooling MRC data per muscle group Patients examined (n)	Before rescoring	After rescoring	Thresholds location		
					Thres- hold 1	Thres- hold 2	Thres- hold 3
Face, Neck and Trunk muscles			Face, Neck and Trunk muscles				
neck flexors	3	336	Normal	Normal	-3.264	1.114	4.919
neck extensors	3	336	Abnormal	Normal	-4.159	-0.561	1.952
sternocleidomastoid Right	2	167	Abnormal	Normal	-7.343	-2.472	0.074
sternocleidomastoid Left	2	167	Abnormal	Normal	-7.356	-2.429	0.013
trapezius Right	2	167	Abnormal	Normal	0.748	1.642	2.459
trapezius Left	2	167	Abnormal	Normal	0.487	1.762	2.452
pectoralis Right	2	167	Abnormal	Normal	-1.001	1.644	4.24
pectoralis Left	2	167	Abnormal	Normal	-1.059	1.475	4.281
masseter Right	1	105	Abnormal	Abnormal			
masseter Left	1	105	Abnormal	Abnormal			
infrapinatus Right	1	105	Abnormal	Normal	-0.448	2.754	6.063
infrapinatus Left	1	105	Abnormal	Normal	-0.169	2.354	6.142
rhomboid Right	1	105	Abnormal	Normal	0.001	3.012	7.695
rhomboid Left	1	105	Abnormal	Normal	0.028	2.819	7.72
Back extensor muscles	1	105	Abnormal	Normal	1.819	2.858	11.714
latissimus dorsi Right	1	105	Abnormal	Normal	0.637	2.691	11.909
latissimus dorsi Left	1	105	Abnormal	Normal	0.442	3.143	11.904
Abdominal muscles	1	105	Abnormal	Normal	1.894	2.611	11.872
Upper extremity muscles			Upper extremity muscles				
shoulder abductors Right	7	1062	Abnormal	Normal	-3.232	-0.683	2.643
shoulder abductors Left	7	1062	Abnormal	Normal	-2.99	-0.727	2.668
arm exorotation Right	1	62	Abnormal	Normal	-7.252	-0.432	7.684
arm exorotation Left	1	62	Abnormal	Normal	-7.247	-0.442	7.689
shoulder adductors Right	1	62	Abnormal	Normal	-6.99	-0.293	7.283
shoulder adductors Left	1	62	Abnormal	Normal	-6.984	-0.307	7.291
elbow flexors Right	7	1062	Abnormal	Normal	-3.454	-0.441	2.672
elbow extensors Right	4	438	Abnormal	Normal	-4.118	-1.027	2.926
elbow extensors Left	4	438	Normal	Normal	-4.038	-1.258	2.796

Table 2. Continued

	Studies Pooling MRC data per		Before	After	Thresholds location		
	(n)	muscle group	rescoring	rescoring	Thres- hold 1	Thres- hold 2	Thres- hold 3
	Patients examined (n)						
wrist extensors Right	7	1062	Abnormal	Normal	-3.015	0.267	2.66
wrist extensors Left	7	1062	Abnormal	Normal	-3.471	0.348	2.685
wrist flexors Right	3	269	Normal	Normal	-5.606	-1.149	2.615
wrist flexors Left	3	269	Abnormal	Normal	-4.333	0.368	2.159
brachioradial Right	1	105	Abnormal	Normal	-3.236	0.691	3.532
brachioradial Left	1	105	Abnormal	Normal	-3.51	-0.128	3.189
pronator Right	1	105	Abnormal	Normal	-3.595	-0.29	6.767
pronator Left	1	105	Abnormal	Normal	-3.521	-0.383	6.771
supinator Right	1	105	Abnormal	Normal	-3.905	0.966	6.373
supinator Left	1	105	Abnormal	Normal	-3.865	0.707	6.399
grip Right	2	167	Abnormal	Normal	-7.509	-4.424	-3.069
grip Left	2	167	Abnormal	Normal	-7.284	-4.55	-2.852
finger extensors Right	3	269	Abnormal	Normal	-2.376	-1.097	3.475
finger extensors Left	3	269	Abnormal	Normal	-2.704	-0.378	3.38
fingers spreaders Right	3	269	Abnormal	Normal	-3.185	0.76	5.329
fingers spreaders Left	3	269	Abnormal	Normal	-3.459	1.21	4.993
finger flexors Right	3	376	Normal	Normal	-5.181	-0.236	5.701
finger flexors Left	3	376	Normal	Normal	-5.462	0.064	5.114
thumb abductor Right	1	102	Abnormal	Normal	0.216	2.122	4.752
thumb abductor Left	1	102	Abnormal	Normal	-0.79	2.448	5.209
thumb adductor Right	1	102	Normal	Normal	-4.886	0.752	3.575
thumb adductor Left	1	102	Abnormal	Normal	-3.942	-0.016	3.38
opponens pollicis Right	1	102	Abnormal	Normal	-4.834	0.631	3.836
opponens pollicis Left	1	102	Abnormal	Normal	-4.623	0.423	3.89
<b>Lower extremity muscles</b>			<b>Lower extremity muscles</b>				
gluteus Right	2	167	Normal	Normal	-2.21	1.781	5.175
gluteus Left	2	167	Normal	Normal	-2.2	1.703	5.337
hip flexors Right	7	1062	Normal	Normal	-2.394	0.152	3.055
hip flexors Left	7	1062	Normal	Normal	-2.295	0.076	3.045
hip abductors Right	2	167	Normal	Normal	-2.295	0.681	4.134
hip abductors Left	2	167	Normal	Normal	-2.299	0.663	4.214
hip adductors Right	2	167	Abnormal	Normal	0.038	2.138	5.434
hip adductors Left	2	167	Abnormal	Normal	0.135	2.223	5.418
knee flexors Right	4	438	Normal	Normal	-1.333	0.92	4
knee flexors Left	4	438	Normal	Normal	-1.3	0.795	4.156
knee extensors Right	7	1062	Normal	Normal	-2.721	-0.418	2.064
knee extensors Left	7	1062	Normal	Normal	-2.739	-0.381	2.052
foot dorsal flexors Right	7	1062	Abnormal	Normal	-1.589	1.463	3.013
foot dorsal flexors Left	7	1062	Abnormal	Normal	-1.578	1.408	3.093
foot plantar flexors Right	4	438	Normal	Normal	-4.278	-0.166	3.273
foot plantar flexors Left	4	438	Abnormal	Normal	-4.218	-0.339	3.209
toes extensors Right	2	207	Abnormal	Normal	-2.459	0.21	2.048
toes extensors Left	2	207	Abnormal	Normal	-2.271	-0.003	2.474
toes flexors Right	1	102	Abnormal	Normal	-4.328	0.921	2.41
toes flexors Left	1	102	Abnormal	Normal	-4.194	0.954	2.497

**Legend to Table 2:** MRC = Medical Research Council. A normal threshold ordering of the MRC grades is coded as a light grey box; if abnormal = dark grey box. See also Figure 1 for examples explaining these codes. Threshold location = location of the thresholds of adjacent MRC response options located on the created ruler (and expressed in logits).

- **Step 2 Rescoring MRC categories:**

Equivalent to the findings of Analysis 1 and based on the location seen of the disordered thresholds (mid-categories 2-4), all muscle groups were systematically rescored to a modified MRC with four categories. Table 2 provides the data for the rescored MRC categories (see last four columns). Ordered thresholds were restored for all muscles except the masseter muscle. A modified version of the MRC grading system was created for clinical use with the following grades: 0, paralysis; 1, severe weakness (defined as >50% loss of strength); 2, slight weakness (<50% loss of strength); and 3, normal strength. A 50% cut-off was based on the following: having four modified response options is having three thresholds (three theoretical intersections between adjacent response options: Threshold 1, 2 and 3); half of the distance between Threshold 3 (representing the intersection between modified MRC grades 2 and 3; location 4.3 logits) and threshold 1 (intersection between modified grades 0 and 1; location -2.98) for all 72 muscle groups is located at 0.66 logits  $[-2.98 (\text{location Threshold 1}) + 0.5 \times 7.28 (0.5 \times \text{distances between Threshold 3 and Threshold 1})]$ , which is very close to the mean for Threshold 2 (intersection between the modified Grades 1 and 2): 0.46.

- **Step 3 Item bias examination**

Differential item functioning was also performed on person factors age, gender, and diagnosis (Supplementary Table 3). Item bias was hardly seen on age and on gender. On diagnosis, 33 muscle groups (45.8%) demonstrated differential item functioning (Supplementary Table 3).

**Table 3.** Summary Rasch analyses statistics for the modification of MRC sumscore in patients with GBS and CIDP

Analysis	Item fit residuals		Person fit residuals		Item-trait chi-square interaction		PSI	Unidimensionality independent t-test (95%CI)
	Mean	SD	Mean	SD	DF	p-value		
Initial	0.147	4.626	-0.562	1.749	108	<0.00001	0.94	0.20 (0.183-0.218)
Final	0.341	1.100	-0.316	1.094	55	0.0891	0.91	NA

**Legend to Table 3:** SD= standard deviation; DF= degrees of freedom; PSI = person separation index. In the final analysis, item and person fit residuals are acceptable, whereas Chi-square is non-significant, indicating invariance across the trait. A PSI of 0.91 indicates a reliable internal consistency. NA = not available; after performing split analyses, Rasch Unidimensional Measurement Model does not provide the opportunity to perform unidimensionality testing.

### Analysis 3: MRC sum score Rasch analysis in patients with Guillain-Barré syndrome and CIDP (see Figure 2)

- **Step 0 General description of patients examined and initial findings**

A total of 619 patients from several cohorts [Guillain-Barré syndrome,  $n = 480$ ; CIDP,  $n = 139$ ;  $n = 272$  females (43.9%) and  $n = 347$  males (56.1%)] were available for these analyses.<sup>14-16,45,46</sup> The original MRC summed score failed to meet the model

expectations. Misfit statistical findings for all three statistical parameters were initially seen (Table 3, “initial” analysis).

- **Steps 1 and 2 Thresholds examination and rescoring**

Similar findings were seen here as the above-mentioned analyses. Eight muscle groups had disordered threshold. For uniformity, all 12 muscle groups were rescored to four response options, thereby restoring threshold ordering.

- **Step 3 Local dependency and creating subsets**

The following steps were driven by the strongest misfit seen to the Rasch model, which was found to be the strong local dependency findings of equivalent (right and left) muscle pairs (e.g. shoulder abductors right and left side; Spearman's correlations:  $\rho = 0.676\text{--}0.831$ ). Therefore, six subsets of items were created, by combining the corresponding muscle pairs (left and right) with each other, improving the statistical parameters and resolving local dependency.

- **Step 4 Unidimensionality examination**

Based on the first principal components analysis, two comparison groups of subsets were formed with three positively loaded (arm muscle subsets) versus three negative loaded (leg muscle subsets). The independent *t*-tests between these two groups suggested acceptable unidimensionality [*t*-test (95% confidence interval): 0.065 (0.047-0.082)].

- **Step 5 Item bias examination**

Uniform differential item functioning was demonstrated on person factor “disease type” for all created muscle subsets, except for the elbow flexors subset. Therefore, each subset of muscle pairs was split in order to obtain specific subsets for the Guillain-Barré syndrome and CIDP, separately. After this, the model was free of any item bias and local dependency. All subsets of items, except the ‘foot dorsal flexors for patients with Guillain-Barré syndrome’, demonstrated fit statistics within required limits. The foot dorsal flexors in Guillain-Barré syndrome had a fit residual of +5.845 ( $p=0.000021$ ), which disturbed Rasch model fitting (Table 3, final analysis for complete model fit after removing this item). However, for practical reasons the structure of the MRC sum score (composed by 12 muscles) was maintained, despite having skewed foot dorsal flexors in Guillain-Barré syndrome subset of item. A high person separation index (0.91) was obtained for the final modified MRC sumscore model.

## Discussion

Manual muscle testing has been used for more than seven decades as a measure for monitoring disease progression and response to therapy in various neuromuscular disorders.<sup>14-22,46</sup> The MRC grading system has been most widely used for this purpose.<sup>1,4</sup> This study systematically examined the discriminatory capacity of the MRC grading

system in a broad mixture of patients with neuromuscular illnesses, assessing a large number of muscles using Rasch method. The original six response categories of the MRC grading system failed to differentiate among patients with various degrees of muscle weakness. Three-quarters of all muscles examined demonstrated disordered thresholds, especially in the mid-response categories (options 2-4). The inability of physicians to apply the apparently intuitive and easily applicable MRC grades in a proper way is consistent with reports criticizing the MRC system.<sup>1,7,8,47</sup> The current paper also shows that the observed disordered thresholds were generally independent of factors like physicians' experience, duration of illness, or type of practice (university-versus community-based). The original MRC grading system inconsistencies were also 'cross-validated' throughout the neuromuscular cohorts, as the findings between the individual disease cohorts were quite equivalent.

After systematically rescored all MRC grades to a modified four category response option, the accuracy of the MRC grading system increased by fulfilling ordered thresholds requirements. While this change from six to four response options might intuitively lower the ability to capture functional changes in a patient, from the current evidence, however, keeping the six responses will give a false sense of precision and potentially increase the error in assessment, which may lead to a false sense of clinically meaningful improvement when it may not exist.

The current paper shows the difficulties with the use of summed scores derived from various muscles tested in patients with Guillain-Barré syndrome and CIDP. However, after Rasch modelling, we were able to present a transformed modified MRC 12 muscle groups summed scores for use in future clinical studies in these disorders.<sup>13</sup> The analyses revealed severe misfit of the foot dorsal flexors. However, since Guillain-Barré syndrome and CIDP are length dependent neuropathies, we decided to keep this muscle group in the final model. The presented Rasch-built modified interval MRC sum score is considered a substantial improvement compared to the evaluation of muscle strength using ordinal based scores, which in essence are not suitable for performing adequate statistics. The modified interval MRC sum score for patients with CIDP should, however, be applied with some caution, because only 139 patients were assessed, which is lower than the proposed sample size requirements for a stable model.<sup>42</sup> Also, the responsiveness of the Rasch-built modified interval MRC summed score for patients with Guillain-Barré syndrome and CIDP needs to be demonstrated in longitudinal studies, which is currently being investigated.<sup>48</sup> However, its personal separation index was high, indicating good ability of the modified scale to differentiate between groups of patients with various degrees for muscle weakness. Finally, since the differential item functioning findings on diagnosis (Supplementary Table 3) demonstrate that neuromuscular illnesses may behave differently, it is conceivable that Rasch-built MRC sum scores are needed for specific illnesses like multifocal motor neuropathy and other neuromuscular diseases as well. These efforts should be the focus of future studies. In conclusion, the original MRC manual muscle testing grading system failed to meet

the Rasch model expectations in various neuromuscular disorders, despite being the standard metric in neurology worldwide. Modification of this grading system to four response categories (0, paralysis; 1, severe weakness; 2, slight weakness; and 3, normal strength) may significantly enhance the ability of clinicians to differentiate degrees of weakness with greater precision and accuracy. Based on this, we have developed a Rasch-built interval MRC summed score for use in future clinical studies evaluating patients with Guillain-Barré and CIDP. Future studies are warranted to improve the solidness of our neurological assessments.

## References

- 1 Dyck, P. J., Boes, C. J., Mulder, D. *et al.* History of standard scoring, notation, and summation of neuromuscular signs. A current survey and recommendation. *J Peripher Nerv Syst* **10**, 158-173 (2005).
- 2 Mitchell, S. W. and Lewis, M. J. Tendon-jerk and muscle-jerk in disease and especially in posterior sclerosis. *Am J Med Sci* **92**, 363-372 (1886).
- 3 Medical, R. C. Medical Research Council. Aids to the investigation of the peripheral nervous system. 1-2 (Her Majesty's Stationary Office, London, 1943).
- 4 Compston, A. Aids to the investigation of peripheral nerve injuries. Medical Research Council: Nerve Injuries Research Committee. His Majesty's Stationery Office: 1942; pp. 48 (iii) and 74 figures and 7 diagrams; with aids to the examination of the peripheral nervous system. By Michael O'Brien for the Guarantors of Brain. Saunders Elsevier: 2010; pp. [8] 64 and 94 Figures. *Brain* **133**, 2838-2844 (2010).
- 5 Brandsma, J. W., Schreuders, T. A., Birke, J. A. *et al.* Manual muscle strength testing: intraobserver and interobserver reliabilities for the intrinsic muscles of the hand. *J Hand Ther* **8**, 185-190 (1995).
- 6 Cuthbert, S. C. and Goodheart, G. J., Jr. On the reliability and validity of manual muscle testing: a literature review. *Chiropr Osteopat* **15**, 4 (2007).
- 7 MacAvoy, M. C. and Green, D. P. Critical reappraisal of Medical Research Council muscle testing for elbow flexion. *J Hand Surg Am* **32**, 149-153 (2007).
- 8 Merlini, L. Measuring muscle strength in clinical trials. *Lancet Neurol* **9**, 1146; author reply 1146-1147 (2010).
- 9 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 10 Andrich, D. Category ordering and their utility. *Rasch Meas Transact* **9**, 464-465 (1996).
- 11 Penta, M., Tesio, L., Arnould, C. *et al.* The ABILHAND questionnaire as a measure of manual ability in chronic stroke patients: Rasch-based validation and relationship to upper limb impairment. *Stroke* **32**, 1627-1634 (2001).
- 12 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1960).
- 13 Kleyweg, R. P., van der Meche, F. G., and Schmitz, P. I. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle Nerve* **14**, 1103-1109 (1991).
- 14 van der Meche, F. G. and Schmitz, P. I. A randomized trial comparing intravenous immune globulin and plasma exchange in Guillain-Barre syndrome. Dutch Guillain-Barre Study Group. *N Engl J Med* **326**, 1123-1129 (1992).
- 15 van Koningsveld, R., Schmitz, P. I., Meche, F. G. *et al.* Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barre syndrome: randomised trial. *Lancet* **363**, 192-196 (2004).
- 16 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet Neurol* **7**, 136-144 (2008).
- 17 van der Kooij, A. J., Barth, P. G., Busch, H. F. *et al.* The clinical spectrum of limb girdle muscular dystrophy. A survey in The Netherlands. *Brain* **119** (Pt 5), 1471-1480 (1996).
- 18 de Die-Smulders, C. E., Howeler, C. J., Thijs, C. *et al.* Age and causes of death in adult-onset myotonic dystrophy. *Brain* **121** (Pt 8), 1557-1563 (1998).
- 19 Van den Berg-Vos, R. M., Franssen, H., Wokke, J. H. *et al.* Multifocal motor neuropathy: long-term clinical and electrophysiological assessment of intravenous immunoglobulin maintenance treatment. *Brain* **125**, 1875-1886 (2002).
- 20 Hagemans, M. L., Winkel, L. P., Van Doorn, P. A. *et al.* Clinical manifestation and natural course of late-onset Pompe's disease in 54 Dutch patients. *Brain* **128**, 671-677 (2005).
- 21 Van Asseldonk, J. T., Franssen, H., Van den Berg-Vos, R. M. *et al.* Multifocal motor neuropathy. *Lancet Neurol* **4**, 309-319 (2005).
- 22 Hermans, M. C., Faber, C. G., De Baets, M. H. *et al.* Rasch-built myotonic dystrophy type 1 activity and participation scale (DM1-Activ). *Neuromuscul Disord* **20**, 310-318 (2010).



- 23 Asbury, A. K. and Cornblath, D. R. Assessment of current diagnostic criteria for Guillain-Barre syndrome. *Ann Neurol* **27 Suppl**, S21-24 (1990).
- 24 AANA. Research criteria for diagnosis of chronic inflammatory demyelinating polyneuropathy (CIDP). Report from an Ad Hoc Subcommittee of the American Academy of Neurology AIDS Task Force. *Neurology* **41**, 617-618 (1991).
- 25 Bushby, K. M. and Beckmann, J. S. The limb-girdle muscular dystrophies--proposal for a new nomenclature. *Neuromuscul Disord* **5**, 337-343 (1995).
- 26 Hirschhorn, R. and Reuser, A. J. J. in *The metabolic and molecular basis of inherited disease*. (eds C. R. Scriver, A. L. Beaudet, W. Sly, & D. Valle) 3389-3420 (McGraw-Hill, 2001).
- 27 EFNS/PNS. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of multifocal motor neuropathy. Report of a joint task force of the European Federation of Neurological Societies and the Peripheral Nerve Society. *J Peripher Nerv Syst* **11**, 1-8 (2006).
- 28 Prior, T. W. Technical standards and guidelines for myotonic dystrophy type 1 testing. *Genet Med* **11**, 552-555 (2009).
- 29 Miescher, G. C. and Steck, A. J. Paraproteinaemic neuropathies. *Baillieres Clin Neurol* **5**, 219-232 (1996).
- 30 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-59 (2006).
- 31 Wright, B. D. Common Sense for Measurement. *Rasch Measurement Transactions* **13**, 704 (1999).
- 32 Svensson, E. Guidelines to statistical evaluation of data from rating scales and questionnaires. *J Rehabil Med* **33**, 47-48 (2001).
- 33 Tennant, A., McKenna, S. P., and Hagell, P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* **7 Suppl 1**, S22-26 (2004).
- 34 Pallant, J. F. and Tennant, A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* **46**, 1-18 (2007).
- 35 van Nes SI, V. E., van Doorn PA, Hermans M, Bakkers M, Kuitwaard K, Faber CG, Merkies ISJ. Vol. 76 337-345 (Neurology, 2011).
- 36 Shaw, F., Wright, B., and Linacre, J. M. Disordered steps? *Rasch Measurement Transactions* **6**, 225 (1992).
- 37 Guttman, L. in *Studies in social psychology in World War II* Vol. 4. *Measurement and prediction* (eds SA Stouffer et al) 60-90 (Princeton University Press, 1950).
- 38 Tennant, A., Penta, M., Tesio, L. et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care* **42**, I37-48 (2004).
- 39 Vandervelde, L., Van den Bergh, P. Y., Goemans, N. et al. ACTIVLIM: a Rasch-built measure of activity limitations in children and adults with neuromuscular disorders. *Neuromuscul Disord* **17**, 459-469 (2007).
- 40 Smith, E. V., Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* **3**, 205-231 (2002).
- 41 Linacre, J. M. Sample Size and Item Calibration Stability. *Rasch Measurement Transactions* **7** (1994).
- 42 Linacre, J. M. Optimizing rating scale category effectiveness. *J Appl Meas* **3**, 85-106 (2002).
- 43 Andrich D, L. G., Sheridan BE, Lyne A, Sheridan B. *Rasch Unidimensional Measurement Models (RUMM2020 Version 4.0)*. (2003).
- 44 Bland, J. M. and Altman, D. G. Cronbach's alpha. *BMJ* **314**, 572 (1997).
- 45 The, D. G.-B. s. s. g. Treatment of Guillain-Barre syndrome with high-dose immune globulins combined with methylprednisolone: a pilot study. The Dutch Guillain-Barre Study Group. *Ann Neurol* **35**, 749-752 (1994).
- 46 Merkies, I. S. J. *Evaluation of scales and measurement instruments in immune-mediated polyneuropathies*. Thesis, Erasmus Medical Center, (2001).
- 47 Schreuders, T. A., Selles, R. W., Roebroek, M. E. et al. Strength measurements of the intrinsic hand muscles: a review of the development and evaluation of the Rotterdam intrinsic hand myometer. *J Hand Ther* **19**, 393-401; quiz 402 (2006).
- 48 Liang, M. H. Evaluating measurement responsiveness. *J Rheumatol* **22**, 1191-1192 (1995).

**Supplemental Table 1.** Available MRC data on the following muscle groups per neuromuscular cohort of patients

	INCAT study	Dutch GBS trial 1992	Dutch GBS trial 2004*	ICE CIDP trial	MMN patients	DM1 patients	Pompe's patients	LGMD patients
# patients examined	n=113	n=147	n=250	n=117	n=102	n=169	n=62	n=105
<b>Face, Neck and Trunk muscles</b>								
neck flexors						X	X	X
neck extensors						X	X	X
sternocleidomastoid right							X	X
sternocleidomastoid left							X	X
trapezius right							X	X
trapezius left							X	X
pectoralis right							X	X
pectoralis left							X	X
masseter Right								X
masseter Left								X
infraspinatus Right								X
infraspinatus Left								X
rhomboid Right								X
rhomboid Left								X
Back extensor muscles								X
Latissimus dorsi Right								X
latissimus dorsi Left								X
Abdominal muscles								X
<b>Upper extremity muscles</b>								
shoulder abductors right	X	X	X	X	X	X	X	X
shoulder abductors left	X	X	X	X	X	X	X	X
arm exorotation right							X	
arm exorotation left							X	
shoulder adductors right							X	
shoulder adductors left							X	
elbow flexors right	X	X	X	X	X	X	X	X
elbow flexors left	X	X	X	X	X	X	X	X
elbow extensors right					X	X	X	X
elbow extensors left					X	X	X	X
wrist extensors right	X	X	X	X	X	X	X	X
wrist extensors left	X	X	X	X	X	X	X	X
wrist flexors right					X		X	X
wrist flexors left					X		X	X
brachioradial Right								X
brachioradial Left								X
pronator Right								X
pronator Left								X
supinator Right								X
supinator Left								X
grip right							X	X
grip left							X	X
finger extensors right					X		X	X
finger extensors left					X		X	X
fingers spreaders right					X		X	X
finger spreaders left					X		X	X

**Supplemental Table 1.** Continued

	INCAT study	Dutch GBS trial 1992	Dutch GBS trial 2004*	ICE CIDP trial	MMN patients	DM1 patients	Pompe's patients	LGMD patients
# patients examined	n=113	n=147	n=250	n=117	n=102	n=169	n=62	n=105
finger flexors left					X	X		X
thumb abductor right					X			
thumb abductor left					X			
thumb adductor right					X			
thumb adductor left					X			
opponens pollicis right					X			
opponens pollicis left					X			
<b>Lower extremity muscles</b>								
gluteus right							X	X
gluteus left							X	X
hip flexors right	X	X	X	X	X	X	X	X
hip flexors left	X	X	X	X	X	X	X	X
hip abductors right							X	X
hip abductors left							X	X
hip adductors right							X	X
hip adductors left							X	X
knee flexors right					X	X	X	X
knee flexors left					X	X	X	X
knee extensors right	X	X	X	X	X	X	X	X
knee extensors left	X	X	X	X	X	X	X	X
foot dorsal flexors right	X	X	X	X	X	X	X	X
foot dorsal flexors left	X	X	X	X	X	X	X	X
foot plantar flexors right					X	X	X	X
foot plantar flexors left					X	X	X	X
toes extensors right					X			X
toes extensors left					X			X
toes flexors right					X			
toes flexors left					X			

**Legend to Supplemental Table 1:** INCAT = Inflammatory Neuropathy Cause And Treatment; GBS = Guillain-Barré syndrome; ICE= Immune Globulin Intravenous for Chronic Inflammatory Demyelinating Polyneuropathy; CIDP= chronic inflammatory demyelinating polyradiculoneuropathy; MMN = multifocal motor neuropathy; DM1 = myotonic dystrophy type 1; LGMD = limb-girdle muscular dystrophy. A total of 72 muscle groups were available for analyses. For the INCAT studies, a total of 113 patients (GBS n=83, CIDP n=22, and monoclonal gammopathy related polyneuropathy of undetermined significance n=8) were examined.

**Supplemental Table 2.** Person factors included for the differential item functioning analyses

Studies	# patients examined	# muscle groups examined	Age	Gender	Adult	Person factors				Country	Amount of DIF analyses performed on muscles
						Duration of illness	Type of disease	Institution	Clinician's experience		
INCAT study	113	12	X	X		X	X		X		5 x 12 muscles = 60
Dutch GBS trial 1992	147	12	X	X	X			X			4 x 12 muscles = 48
Dutch GBS trial 2004*	250	12	X	X	X			X		X	5 x 12 muscles = 60
DM1 patients	169	22	X	X		X	X				4 x 22 muscles = 88
MMN patients	102	36	X	X		X					3 x 36 muscles = 108
Pompe's patients	62	44	X	X		X					3 x 44 muscles = 132
LGMD patients	105	60	X	X	X	X	X				5 x 60 muscles = 300
ICE CIDP study	117	12	X	X		X					3 x 12 muscles = 36
Total:											8 person factors; 32 DIF analyses performed in a total of 832 muscle groups

**Legend to Supplemental Table 2:** DM1 = myotonic dystrophy type 1; CIDP= chronic inflammatory demyelinating polyradiculoneuropathy; GBS = Guillain-Barré syndrome; INCAT = Inflammatory Neuropathy Cause And Treatment; LGMD = limb girdle muscular dystrophy; MMN = multifocal motor neuropathy. For the INCAT studies, a total of 113 patients (GBS n=83, CIDP n=22, and MGUSP n=8) were examined. For person factor categorisation see method section.

**Supplemental Table 3:** Differential item functioning results by age, gender, and diagnosis before and after rescoring items as part of second analyses (MRC Rasch analyses after pooling data at the individual muscle level)

Person factors à DIF forms à	Before rescoring (original MRC six categories grading system)						After rescoring (modified MRC four categories grading system)					
	Age			Gender			Diagnosis			Age		
	UD	NUD	UD	UD	NUD	UD	UD	NUD	UD	UD	NUD	UD
<b>Face, Neck and Trunk muscles</b>												
neck flexors (1)	0.544402	0.514452	0.003389	0.131593	0.000000	0.001115	0.520361	0.523382	0.004846	0.046606	0.000000	0.004316
neck extensors (1)	0.676138	0.553470	0.019508	0.271448	0.000004	0.000005	0.822015	0.364352	0.013678	0.102678	0.000001	0.01751
sternocleidomastoid Right (2)	0.265644	0.668854	0.09011	0.564518	0.079647	0.069097	0.203856	0.917144	0.172835	0.199909	0.03013	0.002346
sternocleidomastoid Left (2)	0.439242	0.440581	0.133229	0.691045	0.058657	0.055359	0.364513	0.999999	0.219244	0.169342	0.0229	0.001994
trapezius Right (2)	0.71459	0.111568	0.631653	0.196418	0.000002	0.000000	0.999912	0.119729	0.799652	0.159357	0.000019	0.000000
trapezius Left (2)	0.455187	0.170561	0.421901	0.076626	0.000005	0.000000	0.917532	0.189021	0.648562	0.056985	0.000003	0.000000
pectoralis Right (2)	0.254794	0.999999	0.48119	0.201357	0.780071	0.077329	0.238144	0.999999	0.20368	0.269883	0.825925	0.395773
pectoralis Left (2)	0.122845	0.999999	0.677866	0.116688	0.536221	0.017482	0.079519	0.999999	0.949606	0.262383	0.903374	0.16342
masseter Right (5)	0.598073	0.634355	0.228285	0.210041	NA	NA	0.640548	0.637957	0.255769	0.196544	NA	NA
masseter Left (5)	0.598073	0.634355	0.228285	0.210041	NA	NA	0.640548	0.637957	0.255769	0.196544	NA	NA
infrapinatus Right (5)	0.669817	0.538622	0.047732	0.227285	NA	NA	0.58229	0.562618	0.023066	0.277492	NA	NA
infrapinatus Left (5)	0.841538	0.93002	0.36698	0.439538	NA	NA	0.88819	0.885959	0.214929	0.493945	NA	NA
rhomboid Right (5)	0.961293	0.716755	0.907823	0.30939	NA	NA	0.949906	0.585994	0.975044	0.3662	NA	NA
rhomboid Left (5)	0.873824	0.426278	0.821357	0.303487	NA	NA	0.888209	0.27329	0.759083	0.345001	NA	NA
Back extensor muscles (5)	0.40985	0.26014	0.318594	0.999999	NA	NA	0.316508	0.336618	0.536797	0.999999	NA	NA
Latissimus dorsi Right (5)	0.010932	0.999999	0.088348	0.1147545	NA	NA	0.013936	0.999999	0.170587	0.435763	NA	NA
Latissimus dorsi Left (5)	0.521747	0.967755	0.05947	0.999999	NA	NA	0.521905	0.863629	0.117998	0.669992	NA	NA
Abdominal muscles (5)	0.365546	0.514131	0.01335	0.589907	NA	NA	0.374646	0.546594	0.056718	0.334833	NA	NA
<b>Upper extremity muscles</b>												
shoulder abductors Right (4)	0.342545	0.452979	0.004272	0.501357	0.000037	0.659116	0.498135	0.73536	0.000504	0.345397	0.001556	0.765786
shoulder abductors Left (4)	0.383823	0.134659	0.003046	0.417439	0.000146	0.999816	0.660517	0.606819	0.000618	0.800456	0.001968	0.934556
arm exorotation Right (6)	0.209073	0.629999	0.871627	0.233908	NA	NA	0.185798	0.406083	0.68485	0.900906	NA	NA
arm exorotation Left (6)	0.213904	0.624919	0.840278	0.247988	NA	NA	0.179969	0.416388	0.710076	0.870994	NA	NA
shoulder adductors Right (6)	0.174661	0.35643	0.27252	0.718822	NA	NA	0.189312	0.114043	0.485923	0.216998	NA	NA
shoulder adductors Left (6)	0.172706	0.360056	0.298798	0.704542	NA	NA	0.188928	0.13367	0.454837	0.196466	NA	NA
elbow flexors Right (4)	0.818581	0.049523	0.063695	0.365207	0.008651	0.749761	0.749446	0.035513	0.028151	0.570703	0.036841	0.895773

elbow flexors Left (4)	0.531831	0.057403	0.021331	0.460173	0.000569	0.612955	0.507905	0.446964	0.007952	0.312698	0.0005345	0.343929
elbow extensors Right (3)	0.076680	0.687185	0.581407	0.094192	0.001301	0.012492	0.036937	0.550403	0.861022	0.930867	0.000176	0.054143
elbow extensors Left (3)	0.524155	0.999987	0.864974	0.007693	0.037934	0.000103	0.264596	0.966461	0.791247	0.151183	0.011622	0.002271
wrist extensors Right (4)	0.62287	0.42237	0.000554	0.116356	0.000074	0.06715	0.46202	0.083902	0.000112	0.0021099	0.000014	0.170415
wrist extensors Left (4)	0.321593	0.713652	0.085451	0.058468	0.000425	0.100252	0.401539	0.181075	0.013625	0.13798	0.000106	0.026226
wrist flexors Right (3)	0.000125	0.999122	0.247041	0.002492	0.000000	0.999999	0.000755	0.934633	0.531141	0.038141	0.000001	0.999999
wrist flexors Left (3)	0.025484	0.104796	0.483428	0.477785	0.000000	0.999999	0.526176	0.622731	0.35717	0.491893	0.001961	0.077156
brachioradial Right (5)	0.548995	0.999999	0.065405	0.12024	NA	NA	0.263193	0.999999	0.32014	0.323076	NA	NA
brachioradial Left (5)	0.312985	0.999999	0.628651	0.45751	NA	NA	0.34814	0.999999	0.319509	0.241939	NA	NA
pronator Right (5)	0.611239	0.567715	0.025712	0.186817	NA	NA	0.621735	0.608002	0.035059	0.194675	NA	NA
pronator Left (5)	0.732092	0.650209	0.043843	0.122078	NA	NA	0.777318	0.736947	0.070721	0.130227	NA	NA
supinator Right (5)	0.417874	0.747293	0.565996	0.325332	NA	NA	0.236498	0.762924	0.336111	0.379233	NA	NA
supinator Left (5)	0.559357	0.614539	0.93131	0.383338	NA	NA	0.283799	0.664639	0.822714	0.41034	NA	NA
grip Right (2)	0.871197	0.627655	0.461121	0.318197	0.450184	0.025484	0.838568	0.961466	0.485411	0.575795	0.957755	0.895641
grip Left (2)	0.336801	0.853464	0.012186	0.311706	0.094389	0.026151	0.545613	0.702907	0.224266	0.18979	0.978638	0.858148
finger extensors Right (3)	0.01512	0.965992	0.391731	0.094953	0.000000	0.010588	0.286727	0.350263	0.479017	0.251648	0.000002	0.999999
finger extensors Left (3)	0.000181	0.931282	0.235919	0.254877	0.000000	0.999999	0.363482	0.774925	0.031445	0.375887	0.000019	0.999999
fingers spreaders Right (3)	0.000123	0.003161	0.862409	0.482137	0.000000	0.999999	0.019913	0.219629	0.126194	0.195233	0.000000	0.728208
finger spreaders Left (3)	0.002529	0.996851	0.574471	0.529545	0.000000	0.999999	0.000608	0.472412	0.912817	0.039686	0.000000	0.999999
finger flexors Right (1)	0.083622	0.343117	0.678175	0.397956	0.000333	0.000000	0.00276	0.999999	0.396337	0.098473	0.000002	0.000000
finger flexors Left (1)	0.086915	0.377535	0.68782	0.376272	0.00035	0.000000	0.305789	0.156477	0.997594	0.049035	0.383056	0.000000
thumb abductor Right (7)	0.686066	0.985606	0.519915	0.25796	NA	NA	0.64081	0.511449	0.217514	0.139029	NA	NA
thumb abductor Left (7)	0.526099	0.925178	0.991924	0.834409	NA	NA	0.40292	0.717485	0.534467	0.580651	NA	NA
thumb adductor Right (7)	0.668097	0.339954	0.592275	0.920028	NA	NA	0.587315	0.739151	0.47304	0.664807	NA	NA
thumb adductor Left (7)	0.180288	0.345583	0.907628	0.482697	NA	NA	0.106995	0.999999	0.656722	0.05527	NA	NA
opponens pollicis Right (7)	0.118028	0.999999	0.246752	0.475221	NA	NA	0.074896	0.999999	0.783657	0.659425	NA	NA
opponens pollicis Left (7)	0.741679	0.391323	0.976227	0.929787	NA	NA	0.489335	0.923066	0.878436	0.834042	NA	NA
<b>Lower extremity muscles</b>												
gluteus Right (2)	0.206094	0.072424	0.019729	0.163913	0.000000	0.000000	0.651743	0.065238	0.113152	0.028224	0.000001	0.000000
gluteus Left (2)	0.294390	0.097740	0.040055	0.138742	0.000000	0.000000	0.757893	0.217652	0.102148	0.022698	0.000000	0.000002
hip flexors Right (4)	0.563528	0.962943	0.014185	0.226237	0.000000	0.306856	0.625816	0.262678	0.038011	0.52114	0.000000	0.225666
hip flexors Left (4)	0.450965	0.707746	0.002646	0.153311	0.000000	0.999999	0.419261	0.002193	0.03293	0.116348	0.000000	0.839165
hip abductors Right (2)	0.273985	0.222313	0.483501	0.340219	0.000000	0.000000	0.278489	0.640328	0.868323	0.487144	0.000002	0.000181
hip abductors Left (2)	0.220408	0.061164	0.543209	0.309795	0.000000	0.000000	0.186327	0.429481	0.719969	0.731231	0.000001	0.000032

Supplemental Table 3: Continued

Person factors à DIF forms à	Before rescoring (original MRC six categories grading system)						After rescoring (modified MRC four categories grading system)					
	Age		Gender		Diagnosis		Age		Gender		Diagnosis	
	UD	NUD	UD	NUD	UD	NUD	UD	NUD	UD	NUD	UD	NUD
<b>Face, Neck and Trunk muscles</b>												
hip adductors Left (2)	0.035820	0.701673	0.961121	0.13312	0.054388	0.868717	0.058804	0.216037	0.899067	0.060633	0.019979	0.999999
knee flexors Right (3)	0.11215	0.354325	0.576346	0.123215	0.000000	0.105563	0.070165	0.838893	0.55649	0.658417	0.000000	0.126054
knee flexors Left (3)	0.096053	0.298437	0.895987	0.642085	0.000000	0.26285	0.036832	0.744384	0.697655	0.360049	0.000000	0.028004
knee extensors Right (4)	0.88723	0.734647	0.29957	0.143069	0.000011	0.999971	0.978357	0.33795	0.327896	0.010658	0.000007	0.999999
knee extensors Left (4)	0.941754	0.455911	0.226349	0.643317	0.000363	0.509558	0.776879	0.015125	0.394267	0.251139	0.000231	0.552333
foot dorsal flexors Right (4)	0.993434	0.747407	0.124342	0.09629	0.000006	0.999999	0.945667	0.23706	0.075722	0.570517	0.000006	0.999999
foot dorsal flexors Left (4)	0.735196	0.199749	0.014567	0.223981	0.000005	0.999999	0.921943	0.16915	0.011079	0.312052	0.000000	0.999999
foot plantar flexors Right (3)	0.274213	0.236925	0.547816	0.975986	0.000000	0.000196	0.310413	0.147848	0.531267	0.597607	0.000000	0.000000
foot plantar flexors Left (3)	0.680336	0.520689	0.586239	0.98430	0.000000	0.000000	0.455835	0.347433	0.82592	0.815241	0.000000	0.000206
toes extensors Right (1)	0.675734	0.613104	0.581405	0.610962	0.648428	0.999999	0.593315	0.505541	0.680629	0.961637	0.452482	0.999999
toes extensors Left (1)	0.422991	0.521318	0.406329	0.999287	0.598135	0.749161	0.66135	0.520996	0.629391	0.999999	0.773308	0.531289
toes flexors Right (7)	0.548493	0.461503	0.142652	0.930147	NA	NA	0.745498	0.157744	0.052003	0.906572	NA	NA
toes flexors Left (7)	0.638358	0.687773	0.019243	0.705204	NA	NA	0.876585	0.20877	0.154085	0.717285	NA	NA

**Legend to Supplemental Table 3:** MRC = Medical Research Council; NA = not applicable. The muscle groups with differential item functioning (DIF) are highlighted in grey.

- 1 Bonferroni adjusted level of  $p < 0.008333$ , expressed as significant level for each Uniform Differential Item Functioning (UD) and Non-uniform Differential Item Functioning (NUD). Data extracted from myotonic dystrophy, Pompe's disease, and limb-girdle muscular dystrophy cohorts.
- 2 Bonferroni adjusted level of  $p < 0.001190$ . Data extracted from Pompe's disease and limb-girdle muscular dystrophy cohorts.
- 3 Bonferroni adjusted level of  $p < 0.002778$ . Data extracted from Pompe's disease, multifocal motor neuropathy, and limb-girdle muscular dystrophy cohorts.
- 4 Bonferroni adjusted level of  $p < 0.001389$ . Data extracted from all cohorts.
- 5 Bonferroni adjusted level of  $p < 0.000235$ . Data from limb-girdle muscular dystrophy cohort.
- 6 Bonferroni adjusted level of  $p < 0.000379$ . Data from Pompe's disease cohort.
- 7 Bonferroni adjusted level of  $p < 0.000463$ . Data from multifocal motor neuropathy cohort.

# Comparing the NIS vs MRC and INCAT sensory scale through Rasch analyses

*Thomas Draak<sup>1</sup>, MD; Els K. Vanhoutte<sup>1</sup>, MD; Sonja I. van Nes<sup>2</sup>, MD, PhD; Kenneth C. Gorson<sup>3</sup>, MD;  
P.A. van Doorn<sup>2</sup>, MD, PhD; David R. Cornblath<sup>4</sup>, MD; Catharina G. Faber<sup>1</sup>, MD, PhD;  
Ingemar S.J. Merkies<sup>1,5</sup>, MD, PhD; on behalf of the PeriNomS study group*

<sup>1</sup>Department of Neurology, University Medical Centre Maastricht, Maastricht, the Netherlands;

<sup>2</sup>Department of Neurology, Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands;

<sup>3</sup>Department of Neurology, St. Elizabeth's Medical Center, Tufts University School of Medicine, Boston, USA; <sup>4</sup>Department of Neurology, Johns Hopkins School of Medicine, Baltimore, USA;

<sup>5</sup>Department of Neurology, Spaarne Hospital, Hoofddorp, the Netherlands.

*Submitted*



## Abstract

We performed a comparison between NIS sensory versus the mISS, and NIS motor versus the MRC sum score in patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and IgM monoclonal gammopathy of undetermined significance related polyneuropathy (MGUSP).

The ordinal data were subjected to Rasch analyses, creating Rasch-transformed (RT)-intervals for all measures. Comparison between measures was based on validity/reliability with an emphasis on responsiveness (using the patient's level of change related to the individually obtained varying standard errors for minimum clinically important difference (MCID-SE)). Eighty stable patients (GBS: 30, CIDP: 30, MGUSP: 20) were assessed twice (entry: 2-observers; 2-4 weeks later: 1-observer), and 137 newly diagnosed or relapsing patients (GBS: 55, CIDP: 59, IgM-MGUSP: 23) were serially examined with 12 months follow-up.

Data modifications were needed to improve model fit for all measures. The sensory and motor scales demonstrated approximately equal and acceptable validity and reliability scores. Responsiveness scores were poor but slightly higher in RT-mISS compared to RT-NISs. Responsiveness was equal for the RT-motor scales, but higher in GBS compared to CIDP; responsiveness was poor in patients with MGUSP, suggesting a longer duration of follow-up in the latter group of patients.

## Introduction

Sensory and motor deficits contribute to limitations in activity and participation in patients with immune-mediated neuropathies.<sup>1-3</sup> The sensory subset of the Neuropathy Impairment Scale (NISs) and the Inflammatory Neuropathy Cause And Treatment (INCAT) sensory scale (ISS) have been widely used to assess sensory deficits.<sup>4-6</sup> The NISs measures touch-pressure, pin-prick, joint position and vibration sense only distally at the index finger and great toe and does not assess any potential proximal sensory loss.<sup>5-8</sup> In addition, vibration is assessed with a non-graduated tuning fork and two-point discrimination at the index finger is not incorporated.<sup>9</sup> The original ISS measures pinprick and vibration sense in the arms and legs from distal to proximal recording the highest extension of sensory abnormality and two-point discrimination at the index finger.<sup>4</sup> Joint position and light touch modalities initially were not incorporated, and therefore this scale was less comprehensive and perhaps less responsive.<sup>7</sup> Additionally, the two-point discrimination categories were arbitrarily assigned.<sup>10</sup> Therefore, we created the modified ISS (mISS) combining the apparent advantages of both scales. The modalities of touch pressure and joint position were incorporated in the mISS, proximal assessment was preserved, and static two-point discrimination values were collected and graded as normal or abnormal based on age-related reference values now available.<sup>11</sup>

The motor subscale of the Neuropathy Impairment Scale (NISm) and the Medical Research Council (MRC) are the two most used grading systems for manual muscle testing.<sup>12,13</sup> Both manual muscle grading systems have been criticized due to the unequal width of the response categories.<sup>14</sup> Although the sensory and motor scales have demonstrated good traditional scientific soundness, it is unknown which scale is superior.<sup>4-6,13,15</sup> Furthermore, these scales are ordinal composite measures, but have been treated as if they are at the interval level with all modalities having equal relevance, which is highly unlikely.<sup>16-20</sup> We investigated whether the NISs, mISS, NISm, and the MRC met Rasch model's expectations, and where needed, measures were taken to transform these ordinal scales into interval scales to meet Rasch's requirements as much as possible. We also compared the scientific properties of each scale between the transformed interval measures with an emphasis on responsiveness to determine whether one outcome measure would be superior to the other.<sup>15,21,22</sup>

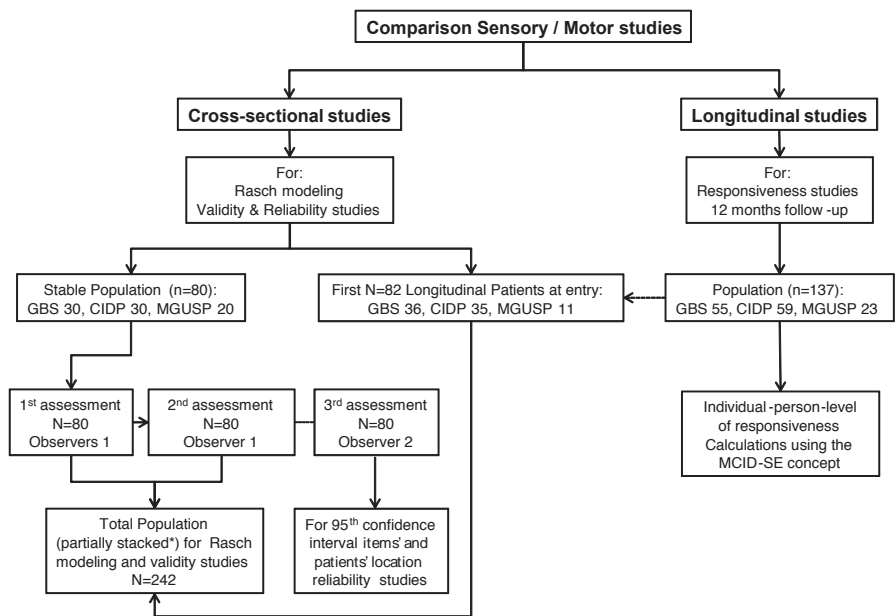
## Methods

### Patients

Cross-sectional group: 80 patients, 30 patients with Guillain-Barré syndrome (GBS), 30 with chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and 20 patients with IgM-monoclonal gammopathy of undetermined significance related

polyneuropathy (MGUSP; all anti-MAG+) with a stable clinical condition were recruited from July 2008 to November 2010. A clinically stable condition was defined as an unchanged functional disability as determined by the patient, and (if applicable) no objective changes of the neurological examination when compared with previous findings in the two months before study enrollment. Patients were recruited from three university outpatient clinics in the Netherlands: Erasmus University Medical Center, University Medical Center Utrecht and Maastricht University Medical Center. Patients were examined twice by the same investigator within an interval of 2-4 weeks, assessing the NISs, mISS, NISm, and the MRC separately without having access to the previous records. During the first visit a second investigator also assessed these scales for inter-rater reliability studies (Figure 1.)

**Figure 1.** Study algorithm



**Legend to Figure 1:** GBS: Guillain-Barré syndrome; CIDP: chronic inflammatory demyelinating polyradiculoneuropathy, MGUSP: monoclonal gammopathy of undetermined significance related polyneuropathy. MCID: minimum clinically important difference; SE: standard error.

Longitudinal group: 137 patients with newly diagnosed or relapsing disease (GBS n=55, CIDP: n=59, and IgM-MGUSP n=23, of which n=19 were anti-MAG+) were enrolled from 9 different countries (26 international centers; Table 1) and were longitudinally examined: GBS/CIDP were assessed at entry, 1, 3, 6, and 12 months, and MGUSP at entry, 3, and 12 months. Patients were treated with intravenous immunoglobulin, plasma exchange, corticosteroids, rituximab, or other immune therapies at the discretion

of their treating physician, and the responsiveness of the scales was assessed in an approach resembling daily clinical practice.

**Table 1.** General characteristics of participants

	Cross-sectional group (for Rasch modeling, validity and reliability studies)	Longitudinal group (for Rasch modeling and responsiveness studies)
Number of patients	80	137
Age (years), mean (SD), range	62 (10), 26-81	54.7 (15.8), 18-90
Gender # (%)		
female	29 (36)	42 (30.7)
male	51 (64)	95 (69.3)
Diagnosis, # (%)		
GBS	30 (38)	55 (40.1)
CIDP	30 (38)	59 (43.1)
MGUSP	20 (25)	23 (16.8)
Country of assessment, # (%)	80	137
the Netherlands	80 (100)	53 (38.7)
USA	-	21 (15.3)
Italy	-	21 (15.3)
Belgium	-	9 (6.6)
Canada	-	16 (11.7)
UK	-	5 (3.6)
Brazil	-	4 (2.9)
France	-	6 (4.4)
Spain	-	2 (1.5)

Eligibility for longitudinally recruited patients included normal cognitive function, age  $\geq 18$  years, fulfilling international criteria for the diagnosis GBS, CIDP, or IgM (preferably positive anti-MAG antibodies)-MGUSP, having clinical deterioration as judged by the patient and physician, and no treatment with immune therapies for at least 2 months prior to inclusion.<sup>23,24</sup> Patients were excluded if other causes of polyneuropathy (e.g. renal insufficiency, diabetes mellitus, etc.) or any other conditions (e.g. alcohol abuse:  $>5$  IU/day) that might interfere with nervous system and physical functioning were present (e.g. arthritis).

All patients gave written informed consent before participation in the study. The local Medical Ethics Committee in each participating neuromuscular center approved the protocol for both cross-sectional and longitudinal studies.

### Outcome measures

Neuropathy Impairment Score – sensory subset (NISs) and Modified INCAT sensory scale (mISS): According to published procedures, touch-pressure was assessed with long-fibre cotton wool, pinprick with straight pins and vibration sense with a non-graduated 165 Hz tuning fork on the dorsal surface of the terminal phalanx of the index finger and great toe at both sides. Furthermore, joint position was assessed by moving the terminal phalanx of the index finger and great toe at both sides.<sup>5,6,8</sup> Each modality

was scored as being normal (0), decreased (1), or absent (2). The NISs ranges from 0 (no sensory deficit) to 32 (most severe sensory deficit). The *mISS* embraces light touch, pinprick, vibration sense, joint position and two-point discrimination assessment. Light touch and pinprick modalities were assessed using disposable cotton wools and pins. Vibration sense was assessed using a graduated 64 Hz Rydel Seiffer tuning fork and published age and location normative values.<sup>9</sup> Joint position was assessed in a standardized manner according to previous recommendations.<sup>25,26</sup> In addition, a static two-point discrimination value was collected using a sliding esthesiometer. Two-point discrimination was assessed only at the right index finger, all other modalities were assessed from distal to proximal in the arms and legs, with only the most affected side being recorded.<sup>4</sup> Modalities were graded as normal (grade 0) or disturbed; at the index finger or hallux (grade 1); at the wrist or ankle (grade 2), at the elbow or knee (grade 3), at shoulder or hip joint (grade 4). Two-point discrimination was graded as normal (grade 0) or disturbed (grade 1) based on age-dependent normative values.<sup>11</sup> The *mISS* ranges from 0 (no sensory deficit) to 33 (most severe sensory deficit).

Neuropathy Impairment Score – motor subset (NISm) and the Medical Research Council sum score (MRCss): In *NISm*, manual muscle strength is graded as 0=normal, 1=25% weak, 2=50% weak, 3=75% weak, 3.25=move against gravity, 3.5=movement, gravity eliminated, 3.75=muscle flicker, no movement, 4=paralysis. A total of 6 muscle pair were tested (arm abductors, forearm flexors, wrist extensors, hip flexors, knee extensors, foot dorsal flexors at both sides) and the NIS-motor ranged from 0 (no motor deficit) to 48 (most severe motor deficit).<sup>5,6,8</sup> The examiner scored deficits by what he considered to be normal considering the test, anatomical site, age, gender, weight, and physical fitness. The *MRCss* of the above mentioned 12 muscles was determined using the MRC grades given in full numbers only, with 0=no movement, no contraction; 1=visible contraction without movement, 3=movement against gravity; 4= movement against resistance, but weaker than normal; 5=normal strength.<sup>27</sup> The *MRCss* ranges from 0 (total paralysis) to 60 (maximal strength).

### Statistical analyses - step 1: From ordinal to interval measures using Rasch

Measurement of changes in disorders with chronic sequel such as GBS (despite its tendency to improve in most patients), CIDP and MGUSP should be performed preferably by using outcome measures that have a high degree of assessment precision, i.e., at the interval or ratio level of assessment rather than using ordinal-based measures.<sup>28,29</sup> All selected measures in the current study are multi-item composite measures at the ordinal level of assessing outcome. A recent study have addressed the deficiencies of the MRC grading system and we expected similar shortcomings for the NIS scale, as this scale is historically related to the MRC.<sup>12,20</sup> Data at the ordinal level can be ordered or ranked, but the difference between the categories is unknown or unequal and therefore, the numbers allocated to the categories have no numerical value and make comparison of changes in these measures impossible.<sup>18,19,28-30</sup>

Analyses of changes using interval or ratio data are more meaningful since the unit of measurement is constant, and the numerical meaning of the numbers is maintained when they are subjected to statistical analysis.<sup>18,20</sup> Therefore, for a proper comparison, the selected measures (NISs, mISS, MRCss, and NISm) were subjected to Rasch analyses to determine whether model expectations would be met.<sup>21,22,31</sup>

To fulfill the minimum requirement of scale stability in Rasch analyses, data from the first observer of the cross-sectional patients (assessment by the first observer at first and second visit: 2 x 80 records) were stacked (controlling for possible confounding “time factor”) and combined with entry data from the first longitudinally examined patients (n = 82) thus creating a total of 242 records for model examination (Figure 1).<sup>32,33</sup>

Additional person factors were arbitrarily categorized aiming for a logical and/or approximately equal distribution between the constructed categories: age (1: <50, 2: 50-59, 3: 60-69 versus 4: ≥70 years), gender (0: female versus 1: male), illness (1: GBS, 2: CIDP, 3: MGUSP), and country (1: the Netherlands, 2: other countries). Analyses were performed to obtain an interval scale for the four outcome measures separately, attempting to maintain the structure of the scales while meeting the Rasch model expectations, thus creating Rasch-Transformed (RT) measures. Requirements for Rasch modeling have been published previously.<sup>22,34</sup>

## Statistical analyses - step 2: Validity, Reliability, and Responsiveness findings

### • Validity and Reliability studies:

Using the assessment of the first observer at visit 1 for the cross-sectional group, we examined the construct convergent validity by correlation between the obtained interval measures for the RT-NISs, RT-mISS, RT-NISm and RT-MRCss using regression with restricted cubic spline function.<sup>35</sup> Internal reliability was examined by determining the Person Separation Index (PSI) for each of the interval scales constructed. A PSI above 0.7 is considered acceptable, indicating the ability to identify at least two groups of patients.<sup>36</sup> In addition, reliability studies were performed comparing the first and second assessment data of the cross-sectional sample to investigate whether hierarchy of items' weight and patients' ability locations were consistent over time.<sup>37</sup> Validity and reliability were quantified by calculation of the intra-class correlation coefficient using a one-way random effects analysis-of-variance (ANOVA) model.

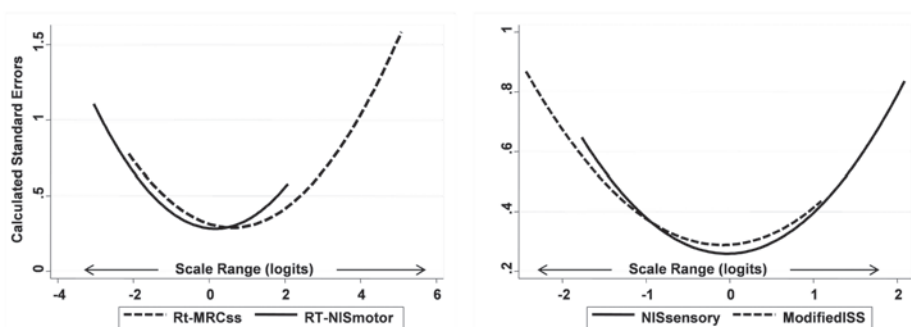
### • Responsiveness studies:

Traditional responsiveness at the *group level* was calculated using the effect size (ES) indicator at 1, 3, 6 and 12 months of follow-up using the traditional sum-score values for the sensory and motor scales separately. The ES is equal to the mean change in score divided by the standard deviation of the scores at entry ( $ES = \mu_i - \mu_o / SD_{\text{entry}}$ ;  $\mu_i$  = mean NISs, mISS, NISm or MRCss of the longitudinally studied group at month =  $i$ ;  $\mu_o$  = mean NISs, mISS, NISm or MRCss at entry).<sup>38</sup> According to Cohen's

rule of thumb, an effect-size value  $\geq 0.8$  represents high responsiveness.<sup>39</sup>

From a modern scientific approach, the Rasch method demonstrates that the error around an individual personal score (standard error (SE)) may vary across the theoretical range of an outcome measure, and with this also the significance of individual changes as shown in supplemental Figure 2.<sup>20,40,41</sup> Therefore, responsiveness at the *individual-person-level* was calculated and referred to as Minimum Clinically Important Difference-Standard Error (MCID-SE) score based on the previously described significant change (SigChange) score.<sup>41</sup> In brief, MCID-SE scores were calculated by computing for each of the 137 longitudinally studied patients separately: a) their individual change (personal-location at  $i$  month minus personal-location at entry; where  $i = 1, 3, 6$ , or 12 months of follow-up for patients with GBS and CIDP and  $i = 3$  or 12 months for patients with MGUSP), and, b) the corresponding SE of difference related to their locations at entry and at  $i$  months ( $SE_{diff} = \text{square-root}(SE_{entry}^2 + SE_{i\text{ month}}^2)$ ), and c) the final MCID-SE calculations by dividing the individual change scores by corresponding  $SE_{diff}$  (MCID-SE = individual change/ $SE_{diff}$ ).<sup>41</sup> As previously suggested, the minimum clinically important difference (MCID) was defined using the cut-off of 1.96 SE (corresponding to a 95% specificity) leading to the following subgroups: subgroup 1 (clinically important improvement (deterioration) for the motor (sensory) scales): MCID-SE  $\geq 1.96$ ; subgroup 2 (clinically unimportant improvement (deterioration)):  $0 < \text{MCID-SE} < 1.96$ ; subgroup 3 (no change): MCID-SE = 0; subgroup 4 (clinically unimportant deterioration (improvement)):  $-1.96 < \text{MCID-SE} < 0$ ; and subgroup 5 (clinically important deterioration (improvement)): MCID-SE  $\leq -1.96$ .<sup>41</sup> Finally, Rasch analyses were performed with the partial credit model as default (RUMM2030).<sup>42</sup> Further statistical analyses were undertaken using Stata 12.0 for Windows XP.

**Figure 2.** Graph showing changing standard error with changing patients' ability location on the sensory and motor metrics



**Legend to Figure 2:** A "U"-shape pattern is shown indicating that the standard error changes with the changing patients' ability location on the theoretical range of the Rasch-transformed (RT-)mISS, the RT-NISs, the RT-NISm and the RT-MRCss. NISs = neuropathy impairment scale – sensory subset. mISS = modified INCAT sensory sumscore, NISm = neuropathy impairment scale – motor subset, MRCss = Medical Research Council sum score

## Results

### General aspects:

The general characteristics of participants are presented in Table 1. A male predominance was seen in both the cross-sectional and longitudinal groups.

### Step 1: Creating interval measures for each ordinal-based outcome measure

- Initial Rasch analyses on sensory and motor measures:

All data were subjected to Rasch analyses for each scale individually. In case of model misfit, the aim was to transform the data to meet model's requirements and also maintain the original structure of the measures as much as possible. Misfit to the Rasch model was demonstrated for both sensory and motor measures (Table 2, initial analysis). The person fit residuals showed acceptable fit statistics, whereas the obtained scores for the item fit residuals (particularly the SD values) deviated from the expected value of 1 (mISS: SD 1.697, NISs: 1.501, MRCss 2.021, NISm 2.110). The significant Chi Square probability for all measures demonstrated no invariance of items. A substantial proportion of the t-tests performed fell outside the  $\pm 1.96$  range, indicating multidimensionality (Table 2).

**Table 2.** Summary statistics for sensory (A/B) and motor (C/D) measures subjected to Rasch analyses

Analyses	Item Fit Residuals		Person Fit Residuals		Item-trait Chi-square interaction		PSI	Unidimensionality Indep. t-test (95%-CI)
	Mean	SD	Mean	SD	DF	p-value		
A	RT-modified ISS							
Initial	0.209	1.697	-0.318	1.037	27	0.00001	0.87	0.07 (0.043-0.098)
Final	0.209	0.671	-0.393	0.940	12	0.383	0.80	0.04 (0.014-0.069)
B	RT-NIS – sensory subset							
Initial	-0.456	1.501	-0.361	0.949	48	0.000006	0.94	0.09 (0.063-0.118)
Final	0.113	1.175	-0.237	0.795	12	0.053	0.84	0.029 (0.001-0.056)
C	RT-MRC sum score							
Initial	-0.025	2.021	-0.376	1.144	36	0.000000	0.91	0.16 (0.134-0.189)
Final	0.082	0.953	-0.456	0.983	24	0.05	0.72576	0.029 (0.001-0.056)*
D	RT-NIS – motor subset							
Initial	-0.313	2.110	-0.337	1.152	36	0.000000	0.7992	0.107 (0.08-0.135)
Final	-0.160	1.262	-0.282	0.764	27	0.001883	0.7323	0.017 (0.011-0.044)*

**Legend to Table 2:** CI: confidence interval, DF: degrees of freedom, PSI: person separation index, SD: standard deviation. RT = Rasch-transformed. For the final statistics using the RT-MRC sum score, we temporarily removed the records of the foot dorsal flexors, since these demonstrated statistical misfit having fit residuals  $> 2.5$ ; however, for maintaining the structure of the original scale the muscle group was kept in. \*For the RT-MRC sum score and the RT-NIS motor subset, unidimensionality was tested before any item was split, since RUMM2030 does not provide the opportunity to do this after splitting. NIS: Neuropathy Impairment Scale; ISS: INCAT sensory scale. MRC: Medical Research Council.



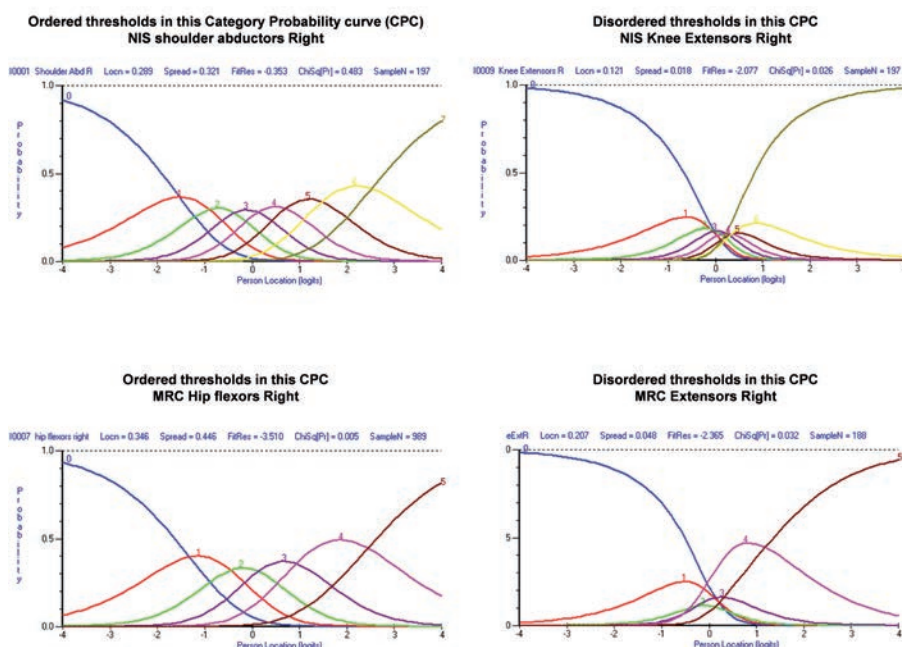
- **Data handling to fit the Rasch model:**

Throughout the analyses, we continuously monitored the distribution of persons within the class intervals, the overall fit statistics, and the independent t-tests for unidimensionality. The 'time factor' introduced as a control for stacking the data showed no confounding factor for both sensory and motor measures.

- **NISs:** All eight sensory modality pairs (e.g., vibration sense on the left and the right great toe, joint position on the left and right index finger) demonstrated a strong local dependency ( $p > 0.696-0.924$ ). Subsequently, eight subsets of items were created combining left and right scores of each modality. Touch pressure and pinprick still demonstrated local dependency for the fingers as well as the toes. Based on these findings and the anatomical distribution of the sensory modalities, a second round of subsets was created combining touch pressure with pinprick, and vibration sense with joint position creating 4 final subsets (somatic sense finger, proprioceptive sense finger, somatic sense toe, proprioceptive sense toe). After these corrections, model's expectations were met, reflected by good fit statistics, no local dependency, and unidimensionality (Table 2, final analysis). There was one uniform item bias (on person factor age category for somatic sense in the legs), which was accepted in order to maintain the internal structure of the NISs. A floor effect was observed in 9.9% of the patients (24/242) and 0.8% had a ceiling effect (2/242).
- **mISS:** Similar to the NISs, correlations were seen between the somatic modalities in the arms and legs, separately. Therefore, similar subsets were created also incorporating two-point discrimination to the subset proprioceptive sense in the arms. There was uniform item bias on person factor age category for somatic sense in the arms and proprioceptive sense in the legs, which was accepted in order to maintain the internal structure of the mISS. The Rasch-Transformed (RT)-mISS fulfilled model's requirements and demonstrated acceptable unidimensionality (Table 2, final analyses). A total of 8.7% (21/242) of the patients demonstrated a floor effect. No ceiling effect was seen.
- **NISm:** Disordered thresholds were seen in all muscle groups except the arm abductors and the forearm flexors (supplemental Figure 3). To maintain uniform response categories, the response options were collapsed for all muscle groups from eight to four categories, taking the frequency distributions into account. After collapsing the response options, the knee extensors still displayed disordered thresholds, but were accepted as part of the model, maintaining the structure of the scale intact. All six muscle pairs (arm abductors, forearm flexors, wrist extensors, hip flexors, knee extensors, foot dorsal flexors) showed high local dependency ( $p > 0.542-0.844$ ). Subsequently, subsets of items were created combining the pairs (left and right) of muscles. Item bias on person factor diagnosis in patients with MGUSP was seen in 3 muscles. Therefore, a subset for these muscles on MGUSP versus others was created. A floor effect was seen in 18.2% of the patients and 1.2% had a ceiling effect.

- **MRCss:** Similar to the NISm, disordered thresholds in response options (supplemental Figure 2), strong local dependency between left and right muscle strength grades, and item bias on diagnosis (MGUSP) were seen. Therefore, response options were reduced to 4 categories (as previously shown),<sup>20</sup> and similar subsets as in the NISm were created. Misfit statistics (fit residual >2.5) were seen in foot dorsiflexors in the subset of GBS and CIDP patients. After removing this, muscle model fit was met. However, to maintain the original scale's structure, we choose not to remove this misfitting muscle (Table 2, final assessment). A floor effect was seen in 0.8% (2/242) and 21.9% had a ceiling effect.

**Figure 3.** Graph showing examples of ordered and disordered findings seen in motor scales



**Legend to Figure 3:** NIS = neuropathy impairment scale. MRC = Medical Research Council sum score.

## Step 2: Clinimetric comparison between the RT-NISs versus RT-mISS, and RT-NISm versus RT-MRCss

### • Validity and reliability studies:

The RT-NISs and RT-mISS demonstrated acceptable correlation with each other (construct validity:  $R^2=0.72$ ;  $p<0.0001$ ). Acceptable PSI scores were obtained for both sensory measures (Table 2, final analyses). Sensory inter-observer reliability of patient's locations was also good (RT-NISs:  $R^2=0.70$ ; RT-mISS:  $R^2=0.74$ ).

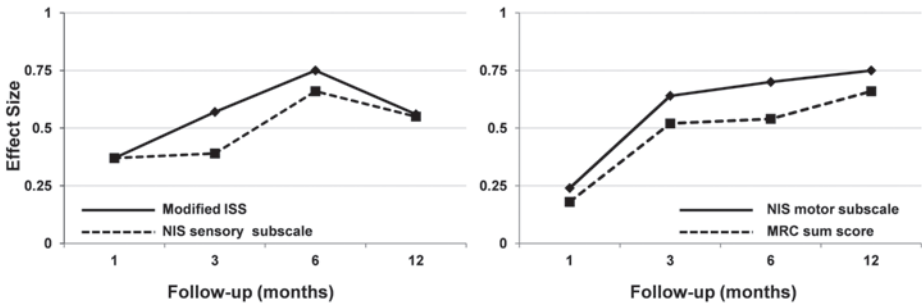
The correlation between the Rasch-transformed motor scales was high ( $R^2=0.93$ ).

Acceptable PSI values were obtained for both motor outcome measures (Table 2). Reliability for the motor scores was moderate to good for the two scales (item's reproducibility: for the RT-NISmotor:  $R^2=0.87$ ; RT-MRCss:  $R^2=0.53$ ; patient's location reproducibility: for the RT-NISmotor:  $R^2=0.58$ ; RT-MRCss:  $R^2=0.67$ ).

• **Responsiveness studies:**

Traditional responsiveness scores using the effect size (ES) varied from poor to moderate for all sensory and motor scales. Slightly better responsiveness scores were seen for the RT-mISS (Figure 4). The ESs scores were higher for the NISmotor versus MRCss.

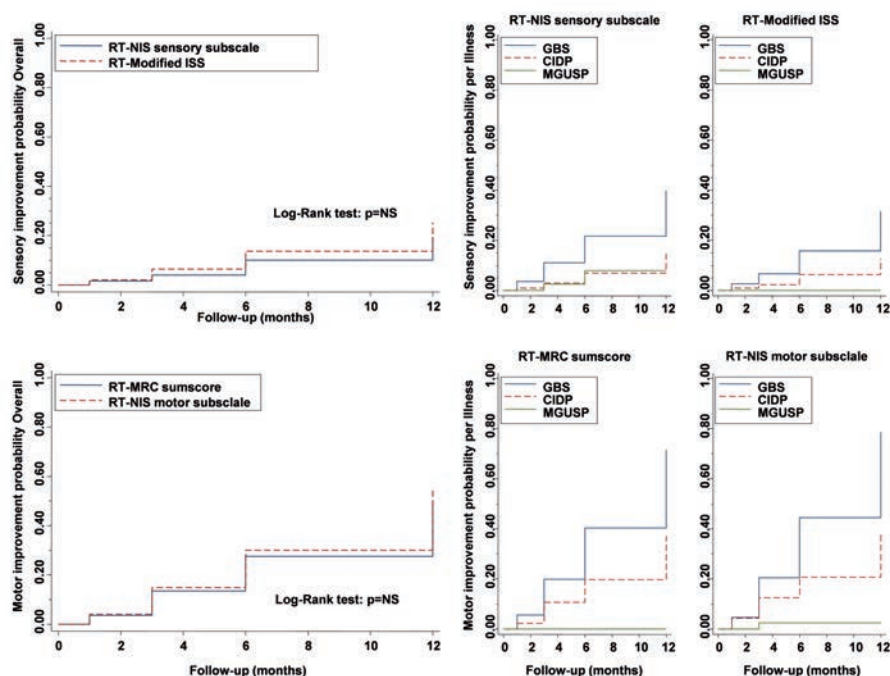
**Figure 4.** Responsiveness findings using the Effect Size in longitudinally examined patients with GBS, CIDP and MGUSP



**Legend to Figure 4:** GBS: Guillain-Barré syndrome; CIDP: chronic inflammatory demyelinating polyradiculoneuropathy, MGUSP: monoclonal gammopathy of undetermined significance related polyneuropathy. ISS: INCAT sensory scale; NIS: Neuropathy Impairment Scale; MRC: Medical Research Council

The responsiveness calculation at the *individual-person-level* taking the changing standard errors into account is shown in Figure 5 (see also supplemental Figure 2). Overall, the sensory responsiveness scores were low, with slightly higher values for the RT-mISS. The scores in the GBS cohort were significantly higher when compared with the CIDP and MGUSP patients. The motor responsiveness findings were higher in both scales compared to the sensory scales and overall comparable with each other. The proportion of patients showing responsiveness was significantly higher in patients with GBS compared to CIDP. In MGUSP, virtually no motor changes were seen (Figure 5). The overall direction and magnitude of clinical changes in the longitudinally examined patients is presented in Table 3, showing the dynamics of shifting patients between the subgroups over time.

**Figure 5.** Responsiveness findings using the MCID-1.96 SE concept in longitudinally examined patients with GBS, CIDP and MGUSP



**Legend to Figure 5:** GBS: Guillain-Barré syndrome; CIDP: chronic inflammatory demyelinating polyradiculoneuropathy, MGUSP: monoclonal gammopathy of undetermined significance related polyneuropathy. ISS: INCAT sensory scale; NIS: Neuropathy Impairment Scale; MRC: Medical Research Council.

## Discussion

This study compared two widely used composite sensory and motor scales in patients with immune-mediated neuropathies.<sup>4-6,13</sup> The scales did not meet the Rasch model expectations showing local dependency between the various sensory modalities, and local dependency and disordered thresholds in the motor scales. After creating subsets, acceptable model fulfillment was accomplished for all measures, thus creating Rasch-Transformed scales at the interval level. These findings clearly demonstrate that composite ordinal measures should not be applied in clinical studies.<sup>18,19</sup> We also found motor > sensory modalities being responsive in GBS and CIDP, which is consistent with other reports.<sup>3,43</sup>

The sensory and motor scales were compared with regard to their clinimetric properties. The RT-mISS and the RT-NISs demonstrated similar clinimetric properties, despite the differences regarding location of assessment (NISs measuring only at the index finger and great toe; in contrast, mISS capturing sensory deficit up to the shoulders and hips).

**Table 3.** Proportion longitudinally examined patients with inflammatory neuropathy categorized using the MCID-SE

Subgroup classification (%) of all longitudinally studied patients based on their MCID-SE values					RT-NIS sensory subset				RT-modified ISS			
Follow-up period (months) à					1	3	6	12	1	3	6	12
1 (clinically important deterioration): MCID-SE $\geq 1.96$					0.9	2.4	1	1	1.9	0	0	0
2 (clinically unimportant deterioration): $0 < \text{MCID-SE} < 1.96$					23.4	20	15.7	17.1	19.4	23.2	15.7	17.1
3 (no change): MCID-SE = 0					13.1	10.4	10.8	13.3	15.7	13.6	10.8	10.5
4 (clinically important improvement): $-1.96 < \text{MCID-SE} < 0$					56.1	60.8	59.8	58.1	54.6	51.2	57.8	59.1
5 (clinically unimportant improvement): MCID-SE $\leq -1.96$					6.5	6.4	12.8	10.5	8.3	12	15.7	13.3
Subgroup classification (%) of all longitudinally studied patients based on their MCID-SE values					RT-NIS motor subset				RT-MRC sum score			
Follow-up period (months) à					1	3	6	12	1	3	6	12
1 (clinically important improvement): MCID-SE $\geq 1.96$					16.4	29.9	36.3	34.9	14.6	27.3	32.7	30.8
2 (clinically unimportant improvement): $0 < \text{MCID-SE} < 1.96$					54.6	41.7	45.1	38.7	52.7	41.4	46.2	46.7
3 (no change): MCID-SE = 0					10.9	16.5	8.8	17.9	13.6	21.1	10.6	14
4 (clinically important deterioration): $-1.96 < \text{MCID-SE} < 0$					15.5	10.2	6.9	7.6	17.3	8.6	7.7	6.5
5 (clinically unimportant deterioration): MCID-SE $\leq -1.96$					2.7	1.6	2.9	0.9	1.8	1.6	2.9	1.9

**Legend to Table 3:** MCID-SE = Minimum Clinically Important Difference-Standard Error score. The MCID-SE score refers to individual-person level responsiveness calculated by individual change divided by their corresponding SE of difference (related to their locations at entry and at  $i$  months where  $i = 1, 3, 6$ , or 12 months of follow-up for patients with GBS and CIDP and  $i = 3$  or 12 months for patients with MGUSP).<sup>41</sup> The highlighted (grey) rows correspond to clinically meaningful improvement. RT: Rasch-transformed. Note: for the sensory scales, a negative MCID-SE score corresponds to an improvement, since both sensory scales range from 0 (no deficit) to 32 and 33 (maximum sensory deficit), for the Neuropathy Impairment scale sensory subset (NISs) and modified INCAT sensory scale (mISS), respectively. MRC: Medical Research Council.

The responsiveness calculations, both the Effect Size and MCID-SE techniques, were poor and comparable between the two sensory scales. For the motor scales, acceptable validity and reliability scores were obtained with higher responsiveness values for GBS when compared to CIDP and MGUSP. The responsiveness scores were insignificantly higher for the NISm, and overall the findings were comparable (see Figure 5).

The Rasch method demonstrated the varying standard errors across the theoretical range of the interval sensory measures, and with this also the significance and direction of clinically important changes in patients with different ability levels (supplemental Figure 2; Table 3).<sup>40,41</sup> Since the direction and magnitude of change is captured using the MCID-SE concept, this design could be applicable in clinical trials.<sup>41</sup> The Rasch method enabled the use of individual standard errors based on patients' ability level. A cut-off value of  $1.96 \times \text{SE}$  has been suggested as this represents a more stringent approach in differentiating between real changes (exceeding  $1.96 \times \text{SE}$ ) compared to "minor non-meaningful" changes. This reduces the potential for imprecise responsiveness. Future studies should focus on determining proper metrics for patients with MGUSP, since none of the scales in the current study captured any meaningful changes in these patients. A longer period of assessment may be needed for this indolent illness, or consensus could be reached on lowering the MCID-SE cut-off to e.g.  $1 \times \text{SE}$ , accepting some imprecise measurements.

In a recent workshop on outcome measures in inflammatory neuropathies, the inflammatory-Rasch-built overall disability scale (inflammatory-RODS) was demonstrated to be more responsive compared to any of the generally applied scales in these conditions, including the sensory and motor scales used in the current study.<sup>31,44</sup> The inflammatory-RODS was chosen by consensus of the meeting participants as the primary outcome for future studies in these conditions.<sup>44</sup> The current findings do not strongly favor one sensory or motor scale over the other. However, based on assessment of both proximal and distal sites, there is a slight preference for the modified ISS. This study also validates recent observations regarding the inability of researchers to differentiate between MRC grades, which was also observed in the NIS motor subset, not being a big surprise, since these two grading systems are historically linked.<sup>12,20</sup>

In conclusion, we constructed interval measures for the NISs, the mISS, the NISm and the MRCss (RT-NISs, RT-mISS, RT-NISm and RT-MRCss) in patients with GBS, CIDP, and MGUSP by subjecting ordinal sensory and motor scales to Rasch analyses. All scales showed comparable validity and reliability aspects. Traditional and modern responsiveness scores were poor in sensory scales and moderate in motor measures, and higher for all measures in GBS compared to CIDP. Rasch-transformed motor and sensory impairment measures should be used in future clinical studies as secondary outcome measures in these conditions, bypassing the substantial deficiencies of classical ordinal-based scales. Longer-term studies are suggested in patients with MGUSP, since hardly any changes were seen over a 1 year period.

## References

- 1 Molenaar, D. S., de Haan, R., and Vermeulen, M. Impairment, disability, or handicap in peripheral neuropathy: analysis of the use of outcome measures in clinical trials in patients with peripheral neuropathies. *Journal of neurology, neurosurgery, and psychiatry* **59**, 165-169 (1995).
- 2 Merkies, I. S., Hughes, R. A., Donofrio, P. *et al.* Understanding the consequences of chronic inflammatory demyelinating polyradiculoneuropathy from impairments to activity and participation restrictions and reduced quality of life: the ICE study. *Journal of the peripheral nervous system : JPNS* **15**, 208-215 (2010).
- 3 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Connecting impairment, disability, and handicap in immune mediated polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **74**, 99-104 (2003).
- 4 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 5 Dyck, P. J., Karnes, J. L., O'Brien, P. C. *et al.* The Rochester Diabetic Neuropathy Study: reassessment of tests and criteria for diagnosis and staged severity. *Neurology* **42**, 1164-1170 (1992).
- 6 Dyck, P. J., Kratz, K. M., Lehman, K. A. *et al.* The Rochester Diabetic Neuropathy Study: design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. *Neurology* **41**, 799-807 (1991).
- 7 Merkies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscular disorders : NMD* **16**, 149-156 (2006).
- 8 Dyck, P. J., Davies, J. L., Litchy, W. J. *et al.* Longitudinal assessment of diabetic polyneuropathy using a composite score in the Rochester Diabetic Neuropathy Study cohort. *Neurology* **49**, 229-239 (1997).
- 9 Martina, I. S., van Koningsveld, R., Schmitz, P. I. *et al.* Measuring vibration threshold with a graduated tuning fork in normal aging and in patients with polyneuropathy. European Inflammatory Neuropathy Cause and Treatment (INCAT) group. *Journal of neurology, neurosurgery, and psychiatry* **65**, 743-747 (1998).
- 10 Dyck, P. J., Hughes, R. A. C., and O'Brien, P. C. in *Peripheral Neuropathy* (eds P.J. Dyck & P.K. Thomas) (WB Saunders Company, 2005).
- 11 van Nes, S. I., Faber, C. G., Hamers, R. M. *et al.* Revising two-point discrimination assessment in normal aging and in patients with polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **79**, 832-834 (2008).
- 12 Dyck, P. J., Boes, C. J., Mulder, D. *et al.* History of standard scoring, notation, and summation of neuromuscular signs. A current survey and recommendation. *Journal of the peripheral nervous system : JPNS* **10**, 158-173 (2005).
- 13 Medical Research Council. *Medical Research Council. Aids to the investigation of the peripheral nervous system.*, (Her Majesty's Stationary Office, 1943).
- 14 van der Ploeg, R. J. and Oosterhuis, H. J. The "make/break test" as a diagnostic tool in functional weakness. *Journal of neurology, neurosurgery, and psychiatry* **54**, 248-251 (1991).
- 15 Streiner, D. L. and Norman, G. R. *Health measurement scales: a practical guide to their development and use.* (Oxford University Press, USA, 2008).
- 16 DeVellis, R. F. Classical test theory. *Medical care* **44**, S50-59 (2006).
- 17 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of clinical epidemiology* **49**, 711-717 (1996).
- 18 Grimby, G., Tennant, A., and Tesio, L. The use of raw scores from ordinal scales: time to end malpractice? *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* **44**, 97-98 (2012).
- 19 Merbitz, C., Morris, J., and Grip, J. C. Ordinal scales and foundations of misinference. *Archives of physical medicine and rehabilitation* **70**, 308-312 (1989).
- 20 Vanhoutte, E. K., Faber, C. G., van Nes, S. I. *et al.* Modifying the Medical Research Council grading system through Rasch analyses. *Brain : a journal of neurology* **135**, 1639-1649 (2012).
- 21 Rasch, G. *Probabilistic models for some intelligence and attainment tests.* (Danmarks Paedagogiske Institut, 1960).

- 22 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 23 Fokke, C., van den Berg, B., Drenthen, J. *et al.* Diagnosis of Guillain-Barre syndrome and validation of Brighton criteria. *Brain: a journal of neurology* **137**, 33-43 (2014).
- 24 EFNS/PNS. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of paraproteinemic demyelinating neuropathies. Report of a Joint Task Force of the European Federation of Neurological Societies and the Peripheral Nerve Society--first revision. *Journal of the peripheral nervous system: JPNS* **15**, 185-195 (2010).
- 25 Gilman, S. Joint position sense and vibration sense: anatomical organisation and assessment. *Journal of neurology, neurosurgery, and psychiatry* **73**, 473-477 (2002).
- 26 Stillman, B. C. *An investigation of the clinical assessment of joint position sense*, The University of Melbourne, (2000).
- 27 Kleyweg, R. P., van der Meche, F. G., and Schmitz, P. I. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle & nerve* **14**, 1103-1109 (1991).
- 28 Hobart, J. C., Cano, S. J., Zajick, J. P. *et al.* Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet neurology* **6**, 1094-1105 (2007).
- 29 Stevens, S. S. On the theory of scales of measurement. *Science* **103**, 677-680 (1946).
- 30 Wright, B. D. and Linacre, J. M. Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation* **70**, 857-860 (1989).
- 31 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 32 Linacre, J. M. Sample size and item calibration stability. *Rasch Measurement Transactions* **7**, 328 (1994).
- 33 Wright, B. Rack and stack: time 1 vs time 2. *Rasch Measurement Transactions* **17**, 905-906 (2003).
- 34 Hermans, M. C., Merkies, I. S., Laberge, L. *et al.* Fatigue and daytime sleepiness scale in myotonic dystrophy type 1. *Muscle & nerve* **47**, 89-95 (2013).
- 35 Herndon, J. E. and Harrell Jr, F. E. The restricted cubic spline hazard model. *Communications in Statistics-Theory and Methods* **19**, 639-663 (1990).
- 36 Fisher, W. P. Reliability statistics. *Rasch Measurement Transactions* **6**, 238 (1992).
- 37 Wright, B. D. and Stone, M. H. *Best Test Design. Rasch Measurement*. (Mesa Press, 1979).
- 38 Kazis, L. E., Anderson, J. J., and Meenan, R. F. Effect sizes for interpreting changes in health status. *Medical care* **27**, S178-189 (1989).
- 39 Cohen, J. *Statistical power analysis for the behavioral sciences*. (Lawrence Erlbaum Associates Inc, 1988).
- 40 Lai, J. and Eton, D. Clinically meaningful gaps. *Rasch Measurement Transactions* **14**, 850 (2002).
- 41 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health technology assessment* **13**, iii, ix-x, 1-177 (2009).
- 42 Rasch models for measurement: RUMM2030 (Perth, 2010).
- 43 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet neurology* **7**, 136-144 (2008).
- 44 Vanhoutte, E. K., Faber, C. G., and Merkies, I. S. 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscular disorders: NMD* **23**, 924-933 (2013).





## **Part III**

# **Activity and participation studies**



# **Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies**

*S.I. van Nes<sup>1</sup>, MD; E.K. Vanhoutte<sup>2</sup>, MD; P.A. van Doorn<sup>1</sup>, MD, PhD; M. Hermans<sup>2</sup>, MD; M. Bakkers<sup>2</sup>;  
K.Kuitwaard<sup>1</sup>, MD; C.G. Faber<sup>2</sup>, MD, PhD; I.S.J. Merkies<sup>1,2,3</sup>, MD, PhD*

<sup>1</sup>Department of neurology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands;

<sup>2</sup>Department of neurology, Maastricht University Medical Center, Maastricht, the Netherlands;

<sup>3</sup>Department of neurology, Spaarne Hospital, Hoofddorp, the Netherlands.

## Abstract

**Objective:** To develop a patient-based, linearly weighted scale that captures activity and social participation limitations in patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP) and gammopathy-related polyneuropathy (MGUSP).

**Methods:** A preliminary Rasch-built Overall Disability Scale (R-ODS) containing 146 activity and participation items was constructed, based on the WHO international classification of Functioning, Disability and Health, literature search and patients' interviews. The preliminary R-ODS was assessed twice (interval: 2-4 weeks; test-retest reliability studies) in 294 patients who experienced GBS in the past (n = 174) or currently have stable CIDP (n = 80) or MGUSP (n = 40). Data were analyzed using the Rasch unidimensional measurement model (RUMM2020).

**Results:** The preliminary R-ODS did not meet the Rasch model expectations. Based on disordered thresholds, misfit statistics, item bias, and local dependency, items were systematically removed to improve the model fit, regularly controlling the class intervals and model statistics. Finally, we succeeded in constructing a 24 item scale that fulfilled all Rasch requirements. "Reading a newspaper/book" and "eating" were the two easiest items; "standing for hours" and "running" were the most difficult ones. Good validity and reliability were obtained.

**Conclusion:** The R-ODS is a linearly weighted scale that specifically captures activity and social participation limitations in patients with GBS, CIDP and MGUSP. Compared to the Overall Disability Sum Score, the R-ODS represents a wider range of item difficulties thereby better targeting patients with different ability levels. If responsive, the R-ODS will be valuable for future clinical trials and follow-up studies in these conditions.

## Introduction

Disability has been proposed as the preferential level for measuring therapeutic response in immune-mediated neuropathies.<sup>1</sup> However, most disability scales used in these disorders are based on the classic test theory (CTT).<sup>2-7</sup> A summary of the disadvantages of CTT has been provided.<sup>2</sup> Generally, patients are requested to complete all items of CTT based scales, even though some may be irrelevant for their level of ability. Furthermore, a sum of item scores is often calculated assuming equal relevance and hence weighting of each item which is highly unlikely.<sup>2,8</sup> Given these limitations, a modern approach like the Rasch technology is needed to develop a scale measuring disability (i.e., activity and social participation limitations) in patients with immune-mediated neuropathies.<sup>9,10</sup> Rasch models the probability that a person will be able to complete an item, only dependent on the item difficulty and the person's level of ability.<sup>9,11,12</sup> It enables scales measuring the same health construct to be equated on the same linear ruler optimizing comparisons of studies, meta-analyses, and systematic reviews.<sup>10</sup>

The AMC Linear Disability Score (ALDS) item bank and ACTIVLIM are modern disability scales.<sup>13-15</sup> However, their calibration may not be representative to patients with immune-mediated neuropathies since these scales are not disease-specific. Therefore, we developed a linearly weighted scale at the activity and participation level, the Rasch-built Overall Disability Scale (R-ODS), specifically for patients with Guillain-Barré syndrome (GBS), Chronic Inflammatory Demyelinating Polyneuropathy (CIDP) and gammopathy related polyneuropathy (MGUSP) and evaluated its validity and reliability aspects.

## Methods

### Participants.

A total of 852 members of the Dutch society of neuromuscular disorders (Vereniging Spierziekten Nederland [VSN]) registered as having a polyneuropathy were initially requested to participate. For this study, we used data from patients who have been diagnosed with GBS, CIDP or MGUSP (254 of the 511 members who returned their questionnaires). Additionally, 40 patients (11 GBS, 22 CIDP and 7 MGUSP) were recruited at the university outpatient clinics of Rotterdam and Maastricht. All participants were recruited between January 2007 and July 2009. Participants aged 18 years and older, with a stable clinical condition, defined as an unchanged physical and functional condition over 2 months prior to the study as declared by the participant, were included. Standard protocol approvals, registrations, and patient consents. The local Medical Ethics Committee approved the protocol. Written informed consent was obtained from all participants.

**Questionnaire development:**

- **Step I:**

International development procedures were applied to construct the preliminary R-ODS.<sup>12,16,17</sup> Since we aimed to develop an outcome measure at the activity and participation level, we critically reviewed all potential items listed under these headings using the International Classification of Functioning, Disability and Health (ICF).<sup>18</sup> The chapters learning and applying knowledge, general tasks and demands, and communication were excluded. We selected defined items from this ICF list with probable relevance to polyneuropathy patients based on known clinical characteristics. Efforts were made to describe these items in a concise, simple and unambiguous way.

- **Step II:**

A systematic Medline and Embase search was performed for literature published over the last 4 decades, reviewing all scales at the activity and participation levels in patients with any form of (poly)neuropathy. Reports published in English were identified with reference tracing using the keywords: GBS, CIDP, acquired/idiopathic (poly)(radiculo)neuropathy, polyneuritis, gammopathy, dysimmune, paraprotein(a) emia, MGUSP, disability (scale), activity limitation, and handicap. Various handbooks of neurological rating scales were also evaluated. Additionally, items were selected from an earlier semi-structured interview focusing on daily activities addressed by patients with inflammatory polyneuropathies.<sup>5</sup> Eventually, we succeeded in constructing the preliminary R-ODS containing 146 items with 3 response options: (0) "impossible to perform", (1) "performed with difficulty", and (2) "easily performed". If an item was not applicable, the patient was requested to answer this item with the option (3) "not applicable".

- **Step III:**

The selected items were judged by 4 neuromuscular experts. Based on the clinical characteristics of GBS, CIDP and MGUSP, items with insufficient face and content validity were removed.<sup>16</sup>

**Procedures:**

Patients received the preliminary R-ODS plus the Overall Disability Sum Score (ODSS) with instructions by mail.<sup>6</sup> They were requested to answer all questions by themselves and to answer "impossible to perform" when unable to complete an item or "able to perform, but with difficulty" when special devices or other forms of assistance were needed. After a period of 2-4 weeks, all participants were again requested to answer all questions of the preliminary R-ODS (test-retest reliability studies)

**Rasch analyses:**

In the model construction, items scored as (3) "not applicable" were interpreted as missing data. Items with > 10% missing values and patients with > 10% unanswered

items were omitted as a quality control procedure. Thereafter, the remaining responses of the preliminary R-ODS were analyzed using Rasch unidimensional measurement methods (RUMM2020), which is based on a mathematical model proposed by Rasch, a Danish mathematician.<sup>19</sup> The finally constructed scale should be unidimensional, free from item bias, and without disordered thresholds or local dependency.<sup>10,20</sup> We checked for possible item bias using the following person factors, categorized arbitrarily as follows: age (<60 years versus ≥60 years), gender (men vs women), diagnosis (GBS vs CIDP vs MGUSP), diagnosis category (acute [GBS] vs chronic [CIDP and MGUSP]), and duration of complaints (<2 years vs 2–5 years versus ≥5 years). Throughout the analyses, we continuously monitored the distribution of persons within the class intervals, the fit statistics and the independent tests for unidimensionality of the scale. Items and persons not fulfilling Rasch model criteria were evaluated and removed one by one if needed. A sample size of approximately 250 is needed to obtain a 99% confidence with a stable item calibration within  $\pm 0.5$  logits, hence providing a stable model.<sup>21</sup> We expected an acceptable degree of precision of the Rasch analyses, since a sample of 294 patients was available.

**Validity:**

Four neuromuscular experts stated that the items of the final R-ODS were ordered as expected. The external construct validity was assessed by correlation with the ODSS (intraclass correlation coefficient after applying quantile regression analysis). To compare the difficulty range of the items of the ODSS and the final R-ODS in relation to the ability range of the patients in this population we used the anchor-based approach in RUMM2020 (targeting validity study). Two identical items in the ODSS and the R-ODS (“turning a key in a lock” and “dressing upper body”) were used as anchors to place both measures and all patients on the same linear ruler.

**Reliability:**

Internal reliability was examined by determining the Person Separation Index (PSI). In general, a PSI above 0.7 is considered as acceptable, indicating the ability to identify at least two groups of patients.<sup>22</sup> Moreover, test-retest reliability studies were performed to investigate whether hierarchy of item difficulty and patient ability location were consistent over time.<sup>23</sup> Reliability was quantified by calculation of the intraclass correlation coefficient using a one-way random effects analysis-of-variance model for group comparison.

**Statistics and software:**

Rasch analyses were performed with the partial credit model as default, using RUMM2020 software. Further analyses were undertaken using Stata 10.0 for Windows XP. The *p* value was adjusted throughout the analyses, based on Bonferroni multiple testing corrections.<sup>24</sup>



## Results

### General description of patients:

A total of 511 of the 852 approached members of the Dutch society of neuromuscular disorders returned the first questionnaires (response rate of 60%). For the current study, we only used data of patients with a clinically stable form of immune-mediated polyneuropathy (174 GBS, 80 CIDP and 40 MGUSP). Most patients (74%) reported walking difficulties, 48% mentioned at least moderate symptoms in their upper limbs. A total of 255 patients completed the second preliminary R-ODS assessment (87%). The basic characteristics of all participants are presented in table 1.

**Table 1.** Basic characteristics of study population (first sample, n=294)

Diagnosis, n (%)	
GBS	174 (59)
CIDP	80 (27)
MGUSP	40 (14)
Age in years; mean (SD), range	60.2 (13.5), 18-91
Gender, n (%)	
Women	132 (45)
Men	162 (55)
Duration of symptoms, years; mean (SD), range	8.2 (7.8), 0.5 - 40
ODSS arm grade, n (%)*	
≤ 1 (normal or minor symptoms, not affecting any)	123 (53.3)
2 (moderate symptoms, not preventing any)	87 (37.7)
3 (severe symptoms, preventing at least one not all)	20 (8.7)
≥ 4 (severe symptoms, preventing all or no purposeful movements)	1 (0.4)
ODSS leg grade, n (%)*	
0 (normal)	59 (25.5)
1-2 (walking affected, with or without gait disturbance)	103 (44.6)
3 (unilateral support)	40 (17.3)
4 (bilateral support)	10 (4.3)
≥ 5 (wheelchair)	19 (8.2)

**Legend to Table 1.** Abbreviations: CIDP = chronic inflammatory demyelinating polyradiculoneuropathy; GBS = Guillain-Barré syndrome; MGUSP = gammopathy-related polyneuropathy; ODSS = Overall Disability Sum Score

\* Corresponding ODSS grades were only collected in 231 of the 294 patients

### Data quality control (Step 0):

Based on experts' opinion, n = 12 items with insufficient face and content validity were removed (questionnaire development step 3). Also, another n = 36 items with >10% missing values and n = 14 patients with >10% unanswered items were omitted as a quality control procedure.

### Initial Rasch analysis on the preliminary R-ODS:

The first Rasch analysis was performed on the remaining 98 items. The mean items' fit residual showed reasonable model fit (mean -0.188, SD 1.119). However, the mean persons fit residual was -0.363 with a standard deviation exceeding the expected value of 1 (SD 1.428). The significant chi-square probability implied no invariance of items. Initially, a proportion of 0.20 (95% CI: 0.18-0.23) of the t-tests performed fell outside the  $\pm 1.96$  range, indicating multidimensionality.

### Stepwise approach to fit data to the Rasch model:

- **Step 1.**  
Disordered thresholds were seen in two items ("driving a car" and "riding a bike"). Apparently, the response categories were conflicting for the patients. These two items were removed.
- **Step 2.**  
Individual item and person fit statistics were inspected. Individual item fit statistics of 19 items demonstrated misfit to the model (fit residuals exceeding  $\pm 2.5$  or a significant chi-square probability or both). These 19 items were removed one by one, starting with the item with the highest deviation. In addition, two patients showed misfit (fit residuals exceeding  $\pm 2.5$ ) and were also removed.
- **Step 3.**  
In six items (e.g. "standing up from squat position", "shaving/epilating", and "vacuum cleaning") men and women with equal ability levels responded in a significant different manner (demonstrating uniform differential item functioning [DIF] on gender). Another three items demonstrated item bias on person factor diagnosis (category) (two items had uniform DIF, one item non-uniform DIF). Finally, one item demonstrated item bias (uniform DIF) on the duration of complaints. All these items were removed one by one.
- **Step 4.**  
A systematic evaluation of the correlation matrix findings (starting with the highest correlations  $\geq 0.7$ , then  $\geq 0.6$ , through to  $\geq 0.28$ ) was performed to identify local dependency (e.g., a person answering positively to the item "able to walk 1 km" will most likely also respond positively on the item "able to walk 100 m"). Item characteristic curves of each correlating item set were subsequently inspected to select the items that best fit the expected model curve. The item showing the most over- or under-discrimination was then removed. Eventually, a total of 43 items were stepwise removed.

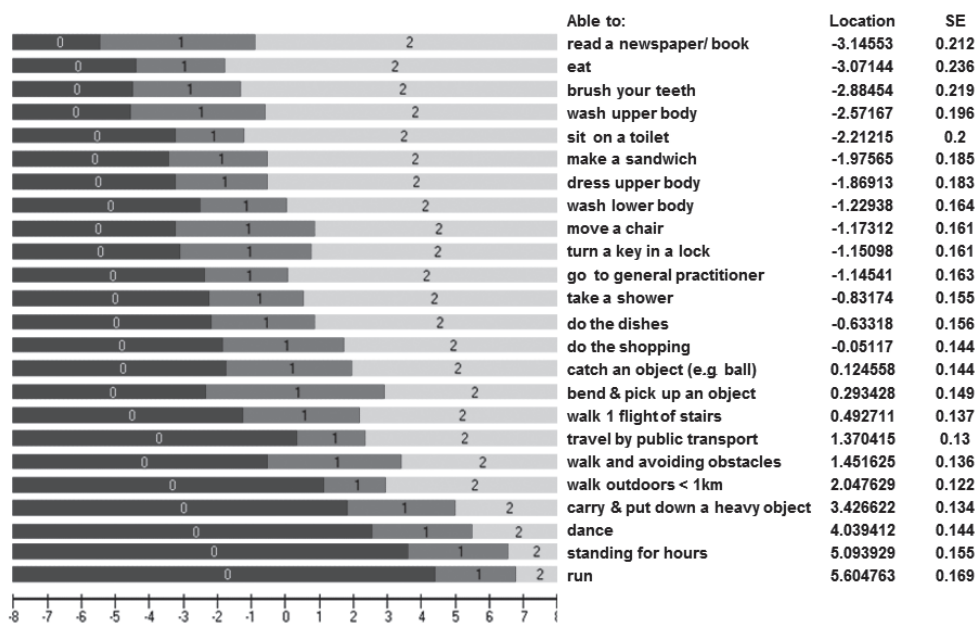
### Final R-ODS

After completing these procedures, we succeeded in constructing a 24 item scale which met the Rasch model expectations, the final R-ODS. Mean fit residuals for items were -0.238 (SD 0.853) and for persons 0.352 (SD 0.839). The chi-square probability

was  $p = 0.29$ , thereby supporting the presence of invariance of item difficulty across the scale. Based on the first principal components analysis two subsets of items were formed with the six most positively loaded vs the six most negatively loaded items. The independent t-tests between these two subsets were significant at the 5% level in 14 of the 261 patients (0.053%, CI 0.027-0.080%) which indicates acceptable unidimensionality.

In the final R-ODS scale, the item “reading a newspaper/book” was the easiest to perform while “running” turned out to be the most difficult task (Figure 1). Item difficulty ranges from -3.15 to 5.60 logits and patient ability level from -5.35 to 8.11 logits. Table 2 provides a nomogram allowing the translation of raw sum scores of the final R-ODS (range 0 to 48) to logits, placing patients’ estimates on the same log-odds units (logit) scale. Since logits are difficult to interpret instinctively, we have converted the person locations into a centile metric score with values ranging from 0 (most severe activity and social participation limitations) to 100 (no activity and social participation limitations).

**Figure 1.** Threshold map and overall item difficulty locations for each items of the final R-ODS



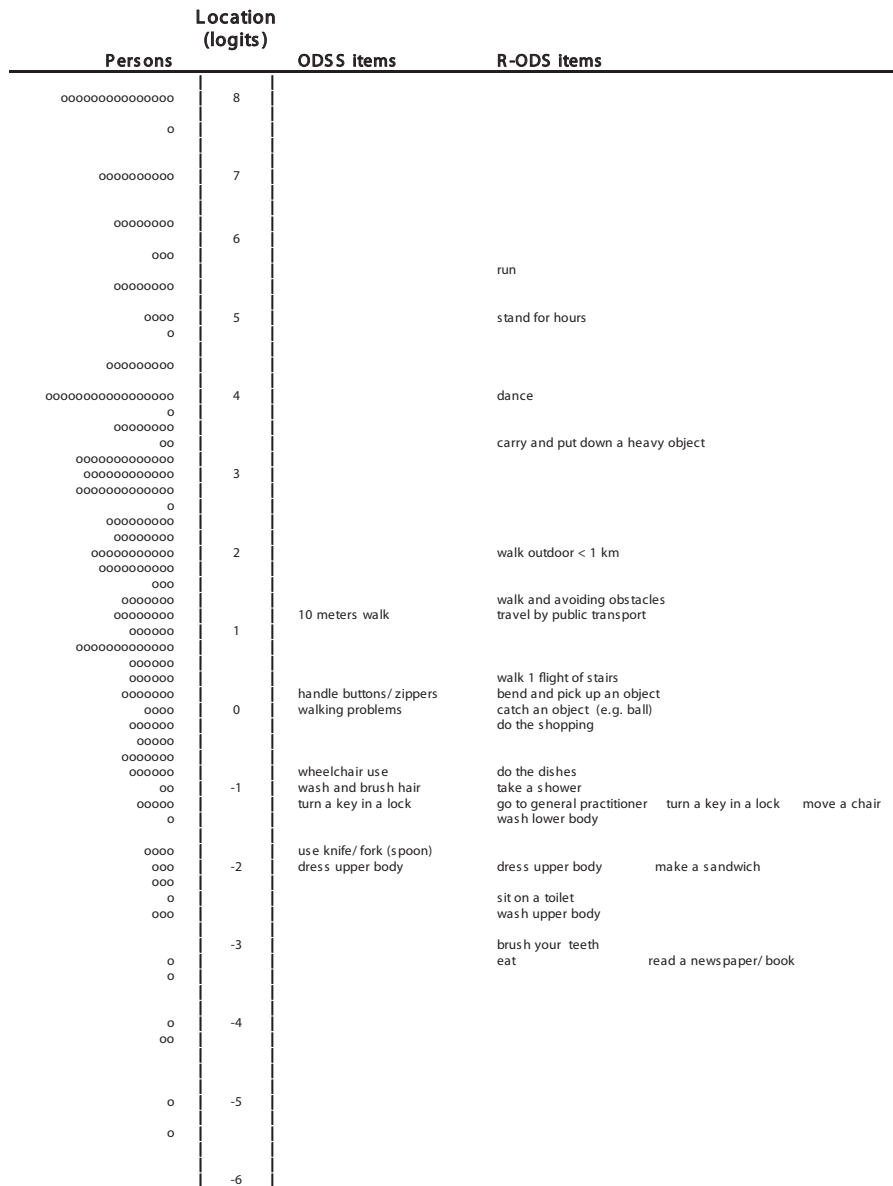
**Legend to Figure 1:** Blue sections (0) = impossible to perform; red sections (1) = with difficulty performed; green sections (2) = easily performed. The length of the colored sections at the left panel represents the number of patients giving a certain response. A gradual shift is seen: the blue section of the item “reading a newspaper/book” is much shorter than the blue section of the item “running”. Patients considered “reading a newspaper/book” the easiest item and “running” the most difficult item. The right panel presents the overall item difficulty locations for each item.

**Table 2.** Nomogram

R-ODS summed raw score	Rasch person location (logits)	Centile metric
0	-6.95	0
1	-6.03	6
2	-5.36	11
3	-4.87	14
4	-4.48	16
5	-4.14	19
6	-3.84	21
7	-3.57	22
8	-3.32	24
9	-3.09	26
10	-2.87	27
11	-2.66	28
12	-2.46	30
13	-2.26	31
14	-2.07	32
15	-1.88	34
16	-1.70	35
17	-1.52	36
18	-1.33	37
19	-1.15	39
20	-0.97	40
21	-0.79	41
22	-0.61	42
23	-0.42	43
24	-0.24	45
25	-0.05	46
26	0.14	47
27	0.34	48
28	0.53	50
29	0.73	51
30	0.94	52
31	1.15	54
32	1.36	55
33	1.58	57
34	1.81	58
35	2.04	60
36	2.28	61
37	2.54	63
38	2.80	65
39	3.09	67
40	3.40	69
41	3.74	71
42	4.11	73
43	4.54	76
44	5.03	80
45	5.59	83
46	6.25	88
47	7.07	93
48	8.11	100

**Legend to Table 2:** Abbreviations: R-ODS = Rasch-built Overall Disability Scale; RUMM2020 = Rasch unidimensional measurement model.

<sup>a</sup> This nomogram allows the translation of raw sum scores of the final R-ODS (range 0–48) to logits or to a centile metric score with values ranging from 0 (most severe activity and social participation restrictions) to 100 (no activity and social participation limitations). The corresponding logits in relation to the raw summed scores is provided by the RUMM software. For convenience, the model also provides an equation for rescoring the logits to more digestible numbers (centile metric).

**Figure 2.** Person locations versus item locations

**Legend to Figure 2:** The numbers 8 to -6 under “Location (logits)” represent corresponding logit scale of a ruler. The open dots under “Persons” represent the number of patients having a person location (ability level) at this point on the ruler. Each open dot corresponds to the location of 1 patient on the ruler. At the right side of the ruler, item locations (item difficulty) of the overall disability sum score (ODSS) and the final R-ODS are presented enabling comparison through targeting (e.g., item difficulty range in relation to the different levels of ability in the sample). GP = general practitioner

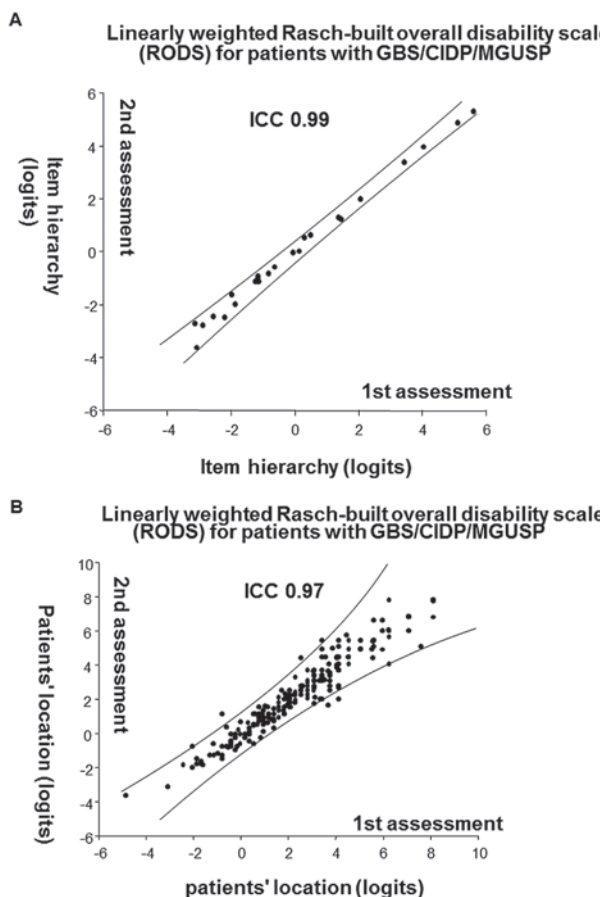
- **Validity**

The intraclass correlation coefficient between the final R-ODS and the ODSS was 0.85, indicating good external construct validity. Furthermore, the population of patients examined demonstrated a ceiling effect of 5.8% on the final R-ODS vs 19.4% ceiling effect on the ordinal ODSS (Student t-test  $p < 0.0001$ ). As shown in Figure 2 item difficulties of the ODSS ranged from -1.869 to 1.336 logits (range: 3.205 logits). By contrast, the final R-ODS demonstrated a wider range of item difficulties (range: 8.750 logits).

- **Reliability**

Internal reliability remained good indicated by a PSI of 0.97 for the final R-ODS. Also, test-retest reliability was good: item hierarchy and patients' location were located within the 95% CI lines, reflecting good invariance (Figure 3A and 3B).

**Figure 3.** Test-retest reliability



**A.** Item difficulty hierarchy of the final R-ODS in the first vs the second assessment. All items (dots) were located within the 95% confidence interval (solid lines) reflecting ideal reliability.

**B.** Patients' location in the first assessment vs the second assessment. Almost all persons (dots) were located within the 95% confidence interval (solid lines). ICC = intraclass correlation coefficient.

## Discussion

We developed the R-ODS, a linearly weighted outcome measure that specifically captures activity and social participation limitations in patients with GBS, CIDP and MGUSP. This questionnaire was constructed using patient's perception of their ability to perform daily and social activities. All 24 items of the final R-ODS met the Rasch model expectations. Although the clinical characteristics of clinically stable GBS, CIDP and MGUSP patients are slightly different, their response pattern was quite similar in this study. Only three items of the preliminary scale were removed due to item bias for the person factor diagnosis.

Since calibration of the R-ODS is based on the characteristics of our study population, we also questioned whether this population represents patients included in clinical trials. The baseline characteristics of patients in trials often report a sum score of the ordinal ODSS of around four. Moreover, inclusion criteria sometimes include minimum sum scores of the ODSS (e.g. at least two in the Randomized Methotrexate CIDP (RMC) trial, evaluating the efficacy of methotrexate in CIDP patients requiring immunoglobulin or corticosteroids) or a minimum ODSS arm or leg grade (e.g. at least arm grade two or leg grade one in a trial evaluating the efficacy of intravenous immunoglobulin in MGUSP).<sup>3,25,26</sup> The basic characteristics of our sample show arm grades  $\geq 2$  in 48% and leg grades  $\geq 1$  in 74% of the patients. This implies sufficient disability in the sample to calibrate the R-ODS for future use in CIDP and MGUSP trials. Only 30% of our sample was unable to walk independently (table 1, ODSS leg grade  $\geq 3$ ) corresponding to a GBS disability score of  $\geq 3$  which is often used as criterion to enroll patients in GBS trials.<sup>27,28</sup> Clearly, patients with GBS in the acute phase may often have more severe symptoms, signs, and functional deficit than our clinically stable patients who experienced GBS in the past. It is therefore conceivable that the weights of the items might have been different if more severely affected patients were included. However, since the PSI of the R-ODS turned out to be extremely high (0.97) it is reasonable to assume that this scale will also be able to differentiate between the various functional phase changes in patients with GBS.<sup>22</sup>

The ALDS item bank and ACTIVLIM are both linearly weighted scales quantifying functional status. The ALDS focusses on chronic conditions and ACTIVLIM on various neuromuscular disorders. We compared the item locations of the R-ODS with the locations of similar items of the ALDS and the ACTIVLIM.<sup>13,15</sup> Not only do the person locations (logits) differ considerably between these scales and the R-ODS, but also the item difficulty ordering differs noticeably. This strengthens the idea that disease-specific scales should be constructed. In our view, although activity and social participation limitations are considered not being disease specific, the significance and weight of corresponding items may appear specific for the various illnesses. Support for this implication is also obtained when looking at the responses of patients having multifocal motor neuropathy (MMN).<sup>29</sup> Our primary aim was to construct an

outcome measure suitable for use in future trials evaluating therapeutic efficacy in a wide range of immune-mediated neuropathies. Therefore, at first not only the current patient population (GBS, CIDP and MGUSP), but also patients with stable MMN were included. However, the primary analyses showed the MMN patients behaving strongly different. Fine motor upper limb items like “turn a key in a lock” and “make a sandwich” were considered much more difficult to accomplish by the MMN group (thus showing significant uniform DIF on person factor “diagnosis”). Their responses to the above items resembled the difficulty levels of items like “walking uphill” and “standing for hours” in our GBS, CIDP and MGUSP patients. We decided to not split the deviating items for the various illnesses. Instead, we are currently constructing a MMN specific activity and participation scale to bypass these limitations.

Both item response theory (IRT) and Rasch measurement estimate the probability that a person with a given ability level will be able to complete an item/task.<sup>9,11,30</sup> The ALDS is based on IRT, which aims to find the best model that explains the data whereas in Rasch observed rating scale data have to satisfy the model expectations.<sup>31-33</sup> As a result, in IRT often a model with generally fewer restrictions is chosen to explain all data. We developed the R-ODS using the Rasch model since it is considered a strong model with many restrictions aiming to satisfy mathematical requirements necessary to achieve fundamental measurement.<sup>34</sup>

There are some limitations that should be addressed. Most participants (86%) were members of a patient organization (VSN). We could not verify their diagnosis, but the board of the VSN has ascertained us that all patients came from centers with great expertise in neuromuscular disorders, ensuring the right diagnosis. Written instructions were given how to respond when assistance or special devices were needed. Nevertheless, there might be some bias of the responses given, depending on the daily and social situation of patients and adaptations made. Furthermore, due to cultural and geographic differences items may be applicable to the Netherlands but not necessarily elsewhere. The ability of the R-ODS to detect relevant clinical changes over time (responsiveness) also needs further evaluation. Its responsiveness and international applicability is currently being investigated in patients with newly diagnosed immune-mediated neuropathies as part of the international multi-center Peripheral Neuropathy Outcome Measures Standardization study.

Nevertheless, the R-ODS seems to be a valid and reliable outcome measure capturing activity and social participation limitations in patients with GBS, CIDP and MGUSP. Compared to the ODSS, the R-ODS represents a wider range of item difficulties, thereby showing a better targeting of the different ability levels of these patients. If its responsiveness can be demonstrated, we expect that the R-ODS will be valuable for future clinical trials and follow-up studies in patients with these disorders.



## References

- 1 Merkies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 2 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-59 (2006).
- 3 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet Neurol* **7**, 136-144 (2008).
- 4 van Swieten, J. C., Koudstaal, P. J., Visser, M. C. *et al.* Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* **19**, 604-607 (1988).
- 5 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 6 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Clinimetric evaluation of a new overall disability scale in immune mediated polyneuropathies. *J Neurol Neurosurg Psychiatry* **72**, 596-601 (2002).
- 7 Graham, R. C. and Hughes, R. A. A modified peripheral neuropathy scale: the Overall Neuropathy Limitations Scale. *J Neurol Neurosurg Psychiatry* **77**, 973-976 (2006).
- 8 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* **49**, 711-717 (1996).
- 9 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1960).
- 10 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 11 Hambleton, R. K., Swaminathan, H., and Rogers, H. J. *Fundamentals of item response theory*. (Sage, 1991).
- 12 Streiner, D. L. and Norman, G. R. *Health measurement scales. A practical guide to their development and use* 2nd ed edn, (Oxford University Press, 1998).
- 13 Holman, R., Weisscher, N., Glas, C. A. *et al.* The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. *Health Qual Life Outcomes* **3**, 83 (2005).
- 14 Holman, R., Lindeboom, R., Vermeulen, M. *et al.* The AMC Linear Disability Score project in a population requiring residential care: psychometric properties. *Health Qual Life Outcomes* **2**, 42 (2004).
- 15 Vandervelde, L., Van den Bergh, P. Y., Goemans, N. *et al.* ACTIVLIM: a Rasch-built measure of activity limitations in children and adults with neuromuscular disorders. *Neuromuscul Disord* **17**, 459-469 (2007).
- 16 Hobart, J. C., Lamping, D. L., and Thompson, A. J. Evaluating neurological outcome measures: the bare essentials. *J Neurol Neurosurg Psychiatry* **60**, 127-130 (1996).
- 17 Bombardier, C. and Tugwell, P. A methodological framework to develop and select indices for clinical trials: statistical and judgmental approaches. *J Rheumatol* **9**, 753-757 (1982).
- 18 WHO. The International Classification of Functioning, Disability and Health. (World Health Organization, Geneva, 2001).
- 19 Rasch Unidimensional Measurement Models (RUMM2020) v. 4.0 (Rumm Laboratory Pty Ltd, Western Australia, 2003).
- 20 Hermans, M. C., Faber, C. G., De Baets, M. H. *et al.* Rasch-built myotonic dystrophy type 1 activity and participation scale (DM1-Activ). *Neuromuscul Disord* **20**, 310-318 (2010).
- 21 Linacre, J. M., Heinemann, A. W., Wright, B. D. *et al.* The structure and stability of the Functional Independence Measure. *Arch Phys Med Rehabil* **75**, 127-132 (1994).
- 22 Fisher, W. P. Reliability statistics. *Rasch Measure Trans* **6**, 238 (1992).
- 23 Wright, B. D. and Stone, M. H. *Best test design*. (Mesa Press, 1979).
- 24 Bland, J. M. and Altman, D. G. Multiple significance tests: the Bonferroni method. *BMJ* **310**, 170 (1995).
- 25 Comi, G., Roveri, L., Swan, A. *et al.* A randomised controlled trial of intravenous immunoglobulin in IgM paraprotein associated demyelinating neuropathy. *J Neurol* **249**, 1370-1377 (2002).

- 26 Randomised controlled trial of methotrexate for chronic inflammatory demyelinating polyradiculoneuropathy (RMC trial): a pilot, multicentre study. *Lancet Neurol* **8**, 158-164 (2009).
- 27 van Koningsveld, R., Schmitz, P. I., Meche, F. G. *et al.* Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barre syndrome: randomised trial. *Lancet* **363**, 192-196 (2004).
- 28 Pritchard, J., Gray, I. A., Idrissova, Z. R. *et al.* A randomized controlled trial of recombinant interferon-beta 1a in Guillain-Barre syndrome. *Neurology* **61**, 1282-1284 (2003).
- 29 Van Asseldonk, J. T., Franssen, H., Van den Berg-Vos, R. M. *et al.* Multifocal motor neuropathy. *Lancet Neurol* **4**, 309-319 (2005).
- 30 Hays, R. D., Morales, L. S., and Reise, S. P. Item response theory and health outcomes measurement in the 21st century. *Med Care* **38**, 1128-42 (2000).
- 31 Massof, R. W. The measurement of vision disability. *Optom Vis Sci* **79**, 516-552 (2002).
- 32 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* **13**, iii, ix-x, 1-177 (2009).
- 33 Tesio, L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J Rehabil Med* **35**, 105-115 (2003).
- 34 Andrich, D. Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med Care* **42**, 17-16 (2004).



# Rasch-built Overall Disability Scale for Multifocal motor neuropathy (MMN-RODS)

*Els K. Vanhoutte<sup>1</sup>, MD; Catharina G. Faber<sup>1</sup>, MD, PhD; Sonja I. van Nes<sup>2</sup>, MD, PhD; Elisabeth A. Cats<sup>3</sup>, MD, PhD; W. Ludo Van der Pol<sup>3</sup>, MD, PhD; Kenneth C. Gorson<sup>4</sup>, MD, PhD; P.A. van Doorn<sup>2</sup>, MD, PhD; David R. Cornblath<sup>5</sup>, MD; Leonard H. van den Berg<sup>3</sup>, MD, PhD; Ingemar S.J. Merkies<sup>1,6</sup>, MD, PhD; on behalf of the PeriNomS study group*

<sup>1</sup>Department of Neurology, University Medical Centre Maastricht, Maastricht, the Netherlands;

<sup>2</sup>Department of Neurology, Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands;

<sup>3</sup>Department of Neurology, Rudolf Magnus Institute of Neuroscience University Medical Centre Utrecht, the Netherlands; <sup>4</sup>Department of Neurology, St. Elizabeth's Medical Center, Tufts University School of Medicine, Boston, USA; <sup>5</sup>Department of Neurology, Johns Hopkins School of Medicine, Baltimore, USA; <sup>6</sup>Department of Neurology, Spaarne Hospital, Hoofddorp, the Netherlands.

*Submitted*

## Abstract

No study has thoroughly examined outcome measures in patients with multifocal motor neuropathy (MMN). Clinical trials in MMN have often used ordinal based measures that may not accurately capture functional changes after treatment.

To construct an interval activity and participation outcome measure specifically for MMN using the Rasch model and to examine its clinimetric properties.

146 Preliminary activity and participation items were assessed twice (reliability studies) in 96 clinically stable MMN patients. These patients also assessed the ordinal-based overall-disability sum-score (ODSS; construct validity). The final Rasch-built overall-disability scale for MMN (MMN-RODS) was serially applied to 26 patients with newly diagnosed or relapsing MMN, treated with intravenous immunoglobulin (1-year follow-up; responsiveness study). The magnitude of change for each patient was calculated using their individual standard errors.

Items (n=121) not fulfilling Rasch requirements were stepwise removed. The final 25-item MMN-RODS fulfilled all Rasch model's expectations and showed acceptable validity, reliability, and responsiveness.

The constructed MMN-RODS is a disease-specific, interval measure that is able to detect limitations in activity and participation in patients with MMN, and overcomes the shortcomings of ordinal scales. Therefore, we recommend its use in future clinical trials in patients with MMN.

## Introduction

Multifocal motor neuropathy (MMN) is an uncommon motor neuropathy characterized by slowly progressive, predominantly distal, asymmetrical limb weakness and persistent conduction block. The diagnosis is mainly based on clinical and electrophysiological findings.<sup>1</sup> Patients with MMN are often treated with intravenous immunoglobulin (IVIg) based upon their beneficial effect in randomized cross-over trials using the Medical Research Council grading system as the primary impairment outcome measure.<sup>1-5</sup> No significant difference was seen at the level of activity and participation in any of these studies.<sup>6</sup> Most studies in MMN have used the (modified) Rankin scale to assess activity and participation.<sup>3-5,7,8</sup> The Guy's disability scale, 10 meter walking test, a self-evaluation scale, a non-validated functional impairment scale, and the Norris scale are disability scales that have also been used in MMN trials.<sup>3,4,9-13</sup> Despite their deficiencies from a modern clinimetric perspective, these activity and social participation outcome measures have been used in MMN in the absence of linearly constructed ability and participation measures.<sup>14,15</sup> They may also be less sensitive to capture relevant clinical changes over time which could have contributed to the non-significant differences observed between patients receiving IVIg compared to those receiving placebo.<sup>6</sup> This study presents the development of an disease-specific interval questionnaire, the Rasch-built overall disability scale for patients with MMN (MMN-RODS) and determines to study its clinimetric properties and thereby its potential value for future research and clinical practice.<sup>13</sup>

## Methods

### Patients

- **Cross-sectional group**

96 patients with MMN, 18 years or older with a stable clinical condition (68 of them received maintenance interval IVIg) were recruited between July 2009 and May 2010. All patients met the clinical and electrophysiological diagnostic criteria for MMN.<sup>16</sup> A clinically stable condition was defined as 'unchanged activities in daily living' from the perspective of the patient, or 'unchanged neurological examination' in the two months preceding the study. Seventy-five patients were recruited by the MMN research team at the Rudolf Magnus institute of neuroscience, University Medical Centre Utrecht in the Netherlands,<sup>17</sup> and 21 were recruited from other MMN centers: 5 from Belgium, 8 from Italy, and 8 from France.

- **Longitudinal group**

Twenty-six patients with MMN (11 newly diagnosed and 15 with clinical relapse) were enrolled from 7 different countries to investigate the responsiveness of the constructed 25-item MMN-RODS (Table 1). All were treated with IVIg.

For both the cross-sectional and the longitudinal group, patients were excluded if concomitant disorders (e.g. diabetes mellitus, renal insufficiency, (prior) chemotherapy, alcohol abuse [ $> 5$  IU/day]) could contribute to pathology or physical functioning. Inclusion and exclusion criteria were part of the Peripheral Neuropathy outcome measures Standardisation (PeriNomS) study, an international collaborative effort of 26 neuromuscular centres worldwide aiming to standardize outcome assessment in inflammatory neuropathies.

**Table 1.** General characteristics of patients with Multifocal Motor Neuropath

	Cross-sectional group (for MMN-RODS scale's construction including validity and reliability studies)	Longitudinal group (for MMN-RODS responsiveness studies)
<b>Number of patients</b>	<b>96</b>	<b>26</b>
Age (years), mean (SD), range	53.3 (11.5), 29 – 86	48.8 (11.8), 29 - 68
<b>Gender, No (%)</b>		
Female	22 (22.9)	5 (19.2)
Male	74 (77.1)	21 (80.8)
<b>New diagnosis of MMN or Relapse</b>		
New diagnosis	-	11 (42.3)
Relapse in known MMN patient	-	15 (57.7)
Duration of symptoms (years), mean (SD), range	12.7 (7.8), 0.2 – 43	2.7 (3.1), 0 – 9.8
<b>Country of assessment, No (%)</b>		
The Netherlands	75 (78.1)	9 (34.6)
Italy	8 (8.3)	5 (19.2)
France	8 (8.3)	3 (11.5)
Belgium	5 (5.2)	1 (3.8)
USA	-	5 (19.2)
Canada	-	1 (3.8)
UK	-	2 (7.7)

### Standard protocol approvals, registrations, and patient consents

The local Medical Ethics Committee in each participating center approved the protocol. Written informed consent was obtained from all participants.

### Questionnaire development

Accepted standardized scale development procedures were applied to create a disease-specific activity and participation scale.<sup>18,19</sup> We critically reviewed potential items listed under the heading “activity and participation” according to the international classification of the consequences of chronic illnesses, after systematic literature review with reference tracing, and after interviewing a focus group of 12 MMN patients.<sup>20</sup> Information elicited from this group was subsequently combined with earlier selected items to form a large item pool, the so-called ‘preliminary Rasch-built overall disability scale’ (preliminary RODS), containing 146 items.<sup>21</sup> Patients were asked to rate their perceived difficulty to perform the selected items on a three-level scale: (0) unable to perform, (1) able to perform, but with difficulty, and (2) easily performed, without difficulty. An item was scored (3) if it was not applicable to the patient.

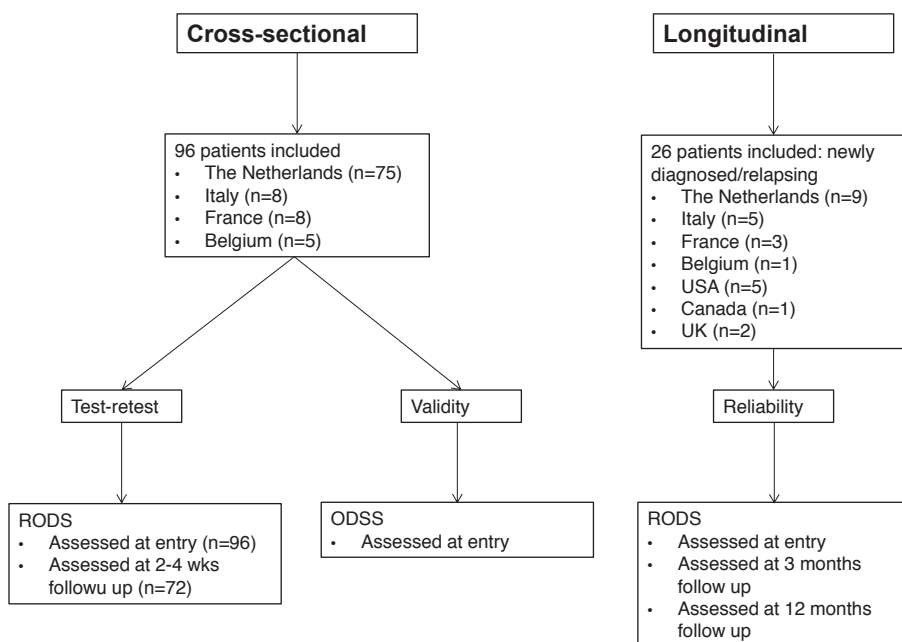
## Additional outcome measure

The overall disability sum score (ODSS) was also applied in the cross-sectional group of patients to assess validity of the final MMN-RODS.<sup>22</sup> The ODSS was previously validated and comprises a functional description of the arms and legs and ranges from 0 ("no signs of disability") to 12 ("most severe disability score"; arm grade: 0–5, leg grade 0–7). The ODSS was chosen since it represents the same level of assessing outcome as the MMN-RODS to be developed.<sup>20</sup> All patients were stable during the study, as declared by their physician and if applicable, based on available clinical records.

## Procedures

After presenting standardized instructions, the cross-sectional group was requested to complete the preliminary RODS questionnaire twice (test-retest study; interval 2–4 weeks; 1<sup>st</sup> assessment: n=96 patients, 2<sup>nd</sup> assessment; n=72 returned the questionnaire). The ODSS was assessed once in the cross-sectional group (validity study). The longitudinal group completed the derived final MMN-RODS at entry, 3, and 12 months of follow-up (Figure 1).

**Figure 1.** Flow chart patient assessment



**Legend to Figure 1:** Flow chart of patient assessment in the cross-sectional and in the longitudinal patient group. ODSS=overall disability sum score. MMN-RODS= Rasch-built Overall Disability Scale for Multifocal Motor Neuropathy.



## Rasch analyses and statistical aspects

- **Rasch description**

The Rasch methodology has been described thoroughly elsewhere.<sup>21,23-26</sup> This statistical technique transforms ordinal obtained scores (which are scale dependent and of limited accuracy) into interval measures that are scale independent and suitably accurate for individual patient assessment.<sup>25</sup> The Rasch method is based on the logical assumption that individuals with greater ability to perform a particular task should, relative to individuals with lower ability levels, have an increased probability of achieving a higher score on the item in question.<sup>24,27</sup> Analyses were performed to obtain a MMN-specific RODS, meeting all Rasch expectations, on a logit (log odds) scale, with proper fit statistical parameters, lack of item bias, no disordered thresholds or local dependency, and the fulfillment of unidimensionality.<sup>21,25,27</sup>

- **Descriptive statistics and Sample size**

Personal factors such as age, gender, duration of symptoms, and country of assessment were collected. These personal factors were categorized as follows: age (1: <40; 2: 40-49; 3: 50-59; 4: ≥60 years), gender (0: female; 1: male), duration of complaints (1: <5; 2: 5-9; 3: 10-20; 4: ≥20 years), and country (1: Holland/Belgium; 2: France; 3: Italy; 4: USA). Age and duration of complaints were categorized striving for ~equal distribution of participants per class interval. According to sample size rules, at least 150 records were needed for acceptable model stability.<sup>28</sup> To fulfill this requirement, the data of the first and second assessments were stacked, controlling for “time factor” as a possible confounding factor.<sup>29</sup> These procedures led to a sample size of 168 patient records (1<sup>st</sup> + 2<sup>nd</sup> assessments) for examination.

- **Reliability and Validity**

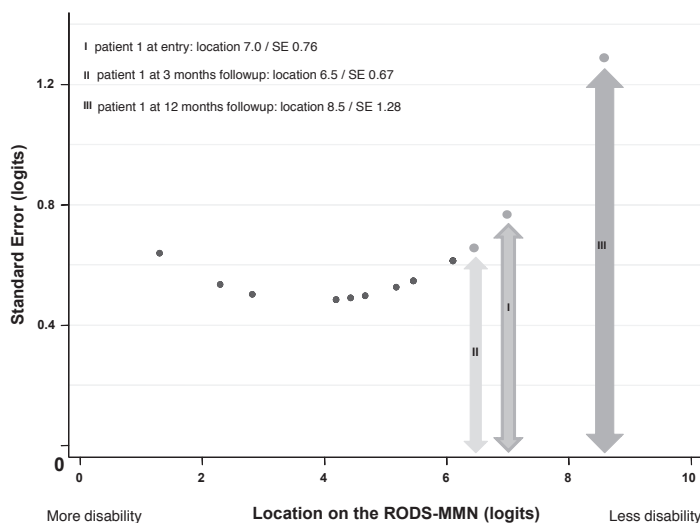
Internal reliability was examined by determining the Person Separation Index (PSI) which indicates the ability to identify at least two groups of patients; a PSI above 0.7 is considered acceptable.<sup>27</sup> Test-retest reliability studies (patients’ ability locations on the logit scale) were also performed to determine the consistency of the final scale created.<sup>30</sup> Reliability was quantified by calculation of the intra-class correlation coefficient using a one-way random effects analysis of variance (ANOVA) model for the comparison of groups. The external construct validity of the final scale was assessed by correlation with the ODSS (ANOVA).<sup>22</sup>

- **Responsiveness**

To define a responder the Minimum Clinically Important Difference (MCID) of an outcome measure is often calculated using a fixed cut-off value based on a fixed estimate of the standard error (SE). However, SE’s may vary across the theoretical range of an outcome measure, with the extreme scores (highest and lowest) having a higher SE than the scores at the centre of an outcome measure. Recognizing the variable SE values to determine the MCID for each individual reveals a dynamic pattern (Figure 2).<sup>31,32</sup> Therefore, responsiveness was calculated at the *individual person level* using Rasch derived individual standard errors (SE) and calculating the

Minimum Clinically Important Difference –Standard Error (MCID-SE). The MCID-SE is based on the previously described concept of significant change (SigChange).<sup>31</sup> The MCID-SE was calculated for each of the 26 serially examined patients separately by computing: a) their own change (personal-location at  $i$  month minus personal-location at entry; where  $i = 3$  or 12 months of follow-up), b) the corresponding SE of difference related to their individual change ( $SE_{diff} = \text{square-root}(SE_{entry}^2 + SE_{i\text{ month}}^2)$ ), and c) the final MCID-SE calculations by dividing the individual change scores by corresponding  $SE_{diff}$  ( $MCID-SE = (\text{personal-location at } i \text{ month minus personal-location at entry})/SE_{diff}$ ).<sup>31</sup> Since MMN is considered an indolent illness, we classified the obtained MCID-SE scores into the following subgroups: subgroup 1 (clinically important improvement):  $MCID-SE \geq 1$  (corresponding to a change of at least 1 x SE); subgroup 2 (clinically unimportant improvement):  $0 < MCID-SE < 1$ ; subgroup 3 (no change):  $MCID-SE = 0$ ; subgroup 4 (clinically unimportant deterioration):  $-1 < MCID-SE < 0$ ; and subgroup 5 (clinically important deterioration):  $MCID-SE \leq -1$ . However, the cut-off of 1 x SE is equivalent to a confidence interval of only 68% certainty; therefore, we also calculated the subgroups using 1.96 X SE cut-off (95% certainty): subgroup 1 (clinically important improvement):  $MCID-SE \geq 1.96$ ; subgroup 2 (clinically unimportant improvement):  $0 < MCID-SE < 1.96$ ; subgroup 3 (no change):  $MCID-SE = 0$ ; subgroup 4 (clinically unimportant deterioration):  $-1.96 < MCID-SE < 0$ ; and subgroup 5 (clinically important deterioration):  $MCID-SE \leq -1.96$

**Figure 2.** Patient location on the MMN-RODS scale with corresponding changes in SE



**Legend to Figure 2:** A "U"-shape pattern is shown indicating that the SE changes with the changing location of a patient on the theoretical range of the MMN-RODS, with the extreme scores (highest and lowest) having a higher SE than the center of an outcome measure. As an example: the red dots correspond with the location of patient 1 at entry (I), at 3 months (II), and 12 months (III) of follow-up. Note the changing corresponding standard error with each time point.

## Software

Rasch analyses were performed with the partial credit model as default (RUMM2030). Further analyses were undertaken using Stata 12.0 for Windows XP. The p-value was adjusted (Bonferroni) throughout the analyses for multiple testing.<sup>33</sup>

## Results

### Study population and data quality control

The study population demographics are presented in Table 1. Most patients were men. In the model construction, items scored as (3) “not applicable” were interpreted as missing data. A total of n=19 items with > 15% missing values were omitted as a quality control procedure.

### Initial Rasch analyses on the preliminary questionnaire

The remaining 127-item questionnaire was subjected to RUMM2030 and showed overall misfit. The items fit residuals showed acceptable fit statistics (mean: -0.360; SD: 1.220), whereas the obtained scores for the person fit residuals deviated from model expectations (mean: -0.428, SD: 1.620). The significant Chi-square probability ( $p < 0.00001$ ) demonstrated no invariance of items and a proportion of 0.20 of the t-tests performed fell outside the  $\pm 1.96$  range, indicating multidimensionality. No differential item functioning (DIF) was seen on ‘time factor’.

### Data handling to fit the Rasch model

Throughout the analyses, we continuously monitored the distribution of persons within the class intervals. In order to improve the model, items and persons not fulfilling Rasch model criteria were step by step evaluated and removed one by one if needed.

- **Step 1:**  
Two items (“get in a car” and “drive a car”) demonstrated disordered thresholds and were removed (n=125 items remaining).
- **Step 2:**  
The individual item fit statistics of 18 items demonstrated misfit to the model (having a significant chi-square probability or having fit residuals exceeding  $\pm 2.5$ ) and were removed one by one (n=107 items remaining).
- **Step 3:**  
Seven items demonstrated item bias or Differential Item Functioning (DIF): 5 items had DIF on person factor country (2 uniform, 2 non-uniform, 1 both), and 2 items demonstrated DIF (1 uniform, 1 non-uniform) on duration of complaints. Also, three items demonstrated DIF (using the 95<sup>th</sup> confidence intervals) on time factor.<sup>30</sup> All 10 items were removed one by one (n=97 items remaining).

- **Step 4:**

Numerous local dependencies were found between many items. All item sets with correlations above 0.30 were evaluated starting with the highest correlations ( $>0.7$ , subsequently  $>0.6$ , etc., to  $>0.30$ ). Of each item set, the item showing least clinical relevance (face and content validity based on consensus between two experts: ISJM and CGF) and with the most over- or under-discrimination on its category probability curve was removed. A total of 72 items were removed in a stepwise manner. After completing these procedures, we succeeded in constructing a 25-item interval measure, the MMN-RODS, that met all Rasch model expectations (item fit residuals: mean  $-0.200$ , SD  $0.993$ ; person fit residuals: mean  $-0.312$ , SD  $1.029$ ; item-trait Chi-square:  $p\text{-value}=0.42$ ,  $DF=50$ ). Two subsets of items were formed, based on the first principal components analysis (the 5 most positively loaded *versus* the 5 most negatively loaded items). The independent t-tests between these two subsets suggested acceptable unidimensionality ( $0.049$  (95%-CI:  $0.013\text{--}0.084$ )). In the final MMN-RODS scale, the item “read a book/newspaper” was the easiest to perform while “button shirt/blouse” turned out to be the most difficult task (Table 2). Item difficulty ranged from  $-2.852$  to  $3.292$  logits. A total of 10 patients demonstrated a maximum score at first assessment (9.3%).

### **Validity, Reliability and Responsiveness studies**

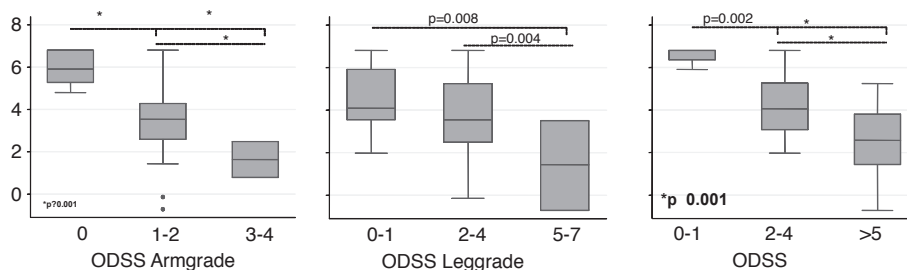
The final 25-item MMN-RODS demonstrated acceptable construct validity when correlated with the ODSS scores (Figure 3). Internal reliability remained robust as indicated by a Person Separation Index of  $0.93$ , indicating great ability to differentiate between patients with MMN with various degrees of activity and participation restrictions in clinical practice.

Figure 4 presents the responsiveness results at the *individual person level* using the MCID-SE  $1 \times SE$  and  $1.96 \times SE$  cut-offs. At 3 months at a 95% confidence interval ( $1.96 \times SE$ ), 10% of patients showed important improvement (subgroup 1), 35% showed unimportant improvement (subgroup 2), 15% remained stable (subgroup 3), 40% unimportant deterioration (subgroup 4) and 0% had important deterioration (subgroup 5). At 12 months, gradual shift was seen for the whole group since 18.8% were categorized in subgroup 1 (important improvement; was 10% at 3 months), 62.5% in subgroup 2 (unimportant improvement; was 35% at 3 months), 12.5% in subgroup 3 (clinically stable; was 15% at 3 months), and 6.3% in subgroup 4 (unimportant deterioration; was 40% at 3 months), and 0% in subgroup 5 (important deterioration) using  $1.96 \times SE$  as a cut off. Of the 26 patients examined serially, 3 patients (11.5%) had no or hardly any activity and participation restrictions.

**Table 2.** Final 25 items Rasch-built overall disability scale for Multifocal Motor Neuropathy (MMN-RODS)

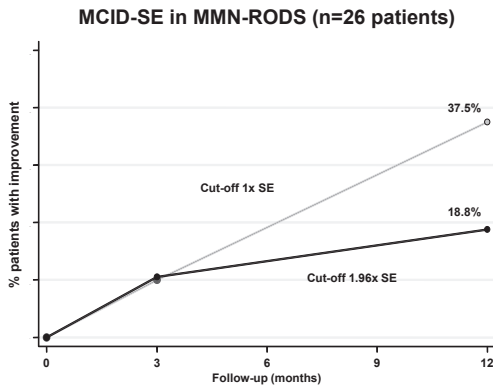
	Are you able to:	unable to perform	able to perform, but with difficulty	able to perform without difficulty
		0	1	2
1	read a book?			
2	make a telephone call?			
3	eat?			
4	open and close a door?			
5	dress your upper body?			
6	brush your teeth?			
7	drink out of mug/glass?			
8	turn a key in a lock?			
9	use knife/fork (spoon)?			
10	clean after toilet?			
11	fill in a form/write?			
12	zip your trousers?			
13	get money from a cash point?			
14	do your own cooking?			
15	pick up small object			
16	work on a computer?			
17	do the bed?			
18	fold laundry?			
19	throw an object (e.g., ball)?			
20	slice vegetables?			
21	peel an apple/orange?			
22	handle small objects (e.g., coin)?			
23	tie your laces?			
24	clip your finger nails?			
25	button your shirt/blouse?			

**Legend Table 2:** The final 25 items MMN-RODS is presented that fulfilled all Rasch model expectations. The easiest item turned out to be 'read a book/newspaper, the most difficult 'button your shirt/blouse. The raw scores can be translated to the less intuitively interpretable logits, and from here to a more understandable centile metric ranging from 0 (most severe activity and participation restrictions) to 100 (no activity limitations and participation restrictions) for the final MMN-RODS scale (data available on request).

**Figure 3.** Construct validity MMN-RODS when compared to the Overall disability sum score (ODSS)

**Legend to Figure 3:** A high patient location on the MMN-RODS refers to minor activity and participation limitations. Low scores on the ODSS also refer to less disability. This figure visualizes the correlations between MMN-RODS patient locations and ODSS scores, reflecting acceptable construct validity. Overall disability sum score (ODSS) ranges from 0 (no disability) to 12 (maximum disability); its armgrade ranges from 0 to 5; its leggrade from 0 to 7.<sup>22</sup>

**Figure 4.** Responsiveness findings in longitudinally examined patients with MMN using two different MCID-SE cut-offs



**Legend to Figure 4:** Percentage of patients classified as important improvement (subgroup 1) during follow-up at 3 and 12 months using two different MCID-SE cut offs. MCID = minimum clinically important difference; SE = Standard Error.

Discussion

The MMN-RODS scale is a Rasch-built interval-weighted outcome measure constructed specifically to capture activity and social participation limitations in MMN patients. This scale was developed noting the limitations seen in most ordinal-based activity and social participation scales, previously known as disability outcome measures, used thus far in MMN.<sup>5,9,12,14,34</sup> The MMN-RODS fulfilled all Rasch model expectations and demonstrated acceptable validity, reliability, and responsiveness at the individual level. These findings contribute to improve the follow-up and monitoring of treatment of patients with chronic conditions and an indolent course like in MMN. After completing the RODS scale, the ordinal obtained raw data need to be transformed using the Rasch model to obtain the ability of the patients in logits with their corresponding SE, thus assessing outcome at the interval level. The PeriNomS study group can be contacted for calculating and interpreting the logits and SE in future studies. In daily practice, the raw scores can be translated to the less intuitively interpretable logits, and from here to a more understandable centile metric ranging from 0 (most severe activity and participation restrictions) to 100 (no activity limitations and participation restrictions) for the final MMN-RODS scale (data available on request).

The clinical presentation of various diseases determines which items or tasks will be completed easily or with more effort. For example, based on differences in the clinical features, it is logical to expect that MMN patients will experience more difficulty with fine motor upper limb items compared to patients with Guillain-Barré syndrome (GBS) being a symmetrical lower limb > upper limb involvement disease. We previously

created a RODS scale specifically for patients with GBS and CIDP.<sup>21</sup> Comparing the RODS questionnaires for GBS/CIDP versus MMN specific for similar and non-disease specific items used for their construction, the locations (weights or difficulties) of items included in the final scale are certainly disease specific and varied considerably. For example, for MMN patients, the item 'button shirt/blouse' (location: 3.292) was the most difficult to accomplish, whereas in GBS/CIDP patients ability to 'run' (location: 5.604 logits) turned out to be the most difficult task to perform.<sup>21</sup> In contrast to the GBS/CIDP-RODS, the MMN-RODS comprises more items related to fine motor activities and less to mobility, which reflects the location of the impairments seen in MMN being predominantly at the distal parts of the upper limbs.<sup>1</sup>

Traditional responsiveness indicators do not always provide information on the personal magnitude and direction (improvement, stable, or deterioration) of changes for each individual patient, as is the case when using modern tools like the Rasch method.<sup>23-25</sup> Hence, traditional methods could be misleading and do not take the changing standard errors observed in individual patients into consideration (Figure 2).<sup>31,32</sup> Relevant clinical changes were clearly captured by the Rasch method, showing improvement at 12 months in most patients, although there was only clinically important improvement in 19% of the patients at a cut-off  $1.96 \times SE$  (Figure 4). MCID is generally defined as a statistic cut-off entity having many faces with no international consensus on which technique to use in general.<sup>35</sup> In addition, the MCID methods applied suggest a static SE, which is highly unlikely when using modern vehicles like IRT and Rasch.<sup>30,31,36</sup> Most studies in MMN having shown 'efficacy' being mainly based on p-value differences or cut-offs arbitrarily taken, which does not necessarily reflect true improvement. Most results seen in MMN studies thus far do not generally take into account the varying imprecise measurements seen when calculating a particular ability of patients. In a recent workshop on outcome measures in immune-mediated neuropathies, consensus was reached to shift towards the use of the MCID-SE concept with varying individually obtained SEs seen across the range of an outcome measure of interest. Clinically important improvement or deterioration should be assessed at the cut-off of  $\pm 1.96 \times SE$  corresponding to a confidence interval of 95% certainty.<sup>37</sup>

IVIg is the first line treatment for MMN.<sup>6,38</sup> In randomized controlled trials performed in MMN various definitions of being a "responder" have been applied.<sup>5,9,12,34,39</sup> However, in these trials, defining 'significant improvement' usually involved only a small increase on a large total score, e.g. in some studies a 1-2 point improvement in the MRC score in 1 or 2 muscles was called "significant" while > 10 muscle or groups were examined. Improvement has also been stated if there was a significant p-value reflecting 'difference', without taking the magnitude or clinical relevance of these changes into account. Since MMN is considered generally an indolent illness with subtle changes, it is not surprising that the observed responsiveness in our study were relatively small over the 12 months follow-up period. However, most serially examined patients showed improvement at 1 year, but only 1/5 was clinically important. Based on these

observations, we tend to suggest future studies in MMN to be conducted over a much longer period aiming to capture more patients having meaningful changes using the proposed MCID-SE cut-offs.

The MMN-RODS is at the interval level which is considered a higher level of measurement precision than ordinal-based scales such as the ODSS.<sup>40</sup> The MMN-RODS also bypasses the general deficiencies known of ordinal measures.<sup>15,41</sup> In addition, using the MMN-RODS through Rasch analyses enables a “patient-targeted approach”, capturing the direction and magnitude of changes seen in each treated individual patient, an approach that can be easily based on the MCID-SE concept proposed in the current paper.

Some methodological limitations in this study need to be addressed. First, In the ideal situation, a sample size of 150, but preferably approximately 250 patients is needed to provide accurate model stability.<sup>28</sup> In the current study, the minimal requirement could only be achieved after stacking the data to a total sample size of 168 records.<sup>29</sup> However, for an orphan disease like MMN this is a large series of patients and no differential item functioning was seen on “time factor”. Second, there was a ceiling effect in 9.3%, which could hamper the applicability of the scale. This was also the case in 3 of the 26 longitudinally examined patients. Hence, selecting patients for future clinical studies should be performed with caution, since not all patients with MMN will demonstrate severe impairment or activity and participation restrictions and will therefore hardly show changes on any scale used. Including these patients in a trial will lower the sensitivity of the chosen metric to capture meaningful changes. Third, the current MMN-RODS is validated for the Netherlands, Belgium, France, and Italy. However, a broader international validation is needed for this metric. Efforts are currently being undertaken by the inflammatory neuropathy consortium to create an international MMN database that could provide more data from various countries, enabling the cross-cultural validation of the created MMN-RODS. Fourth, the MMN-RODS related items mainly reflect upper limb function. It is conceivable that the scale may not adequately estimate the ability in patients with predominantly lower limb involvement. However, its correlation with the leg-grade of the ODSS suggests that, despite not having specific leg functional items, it still captures disability in MMN being leg function related.

In conclusion, the constructed MMN-RODS fulfills all Rasch model requirements and substantially increases our ability to accurately measure changes over time at the level of daily and social activities. Therefore, we recommend its use in future clinical trials in patients with MMN.



## References

- 1 Van Asseldonk, J. T., Franssen, H., Van den Berg-Vos, R. M. *et al.* Multifocal motor neuropathy. *Lancet Neurol* **4**, 309-319 (2005).
- 2 Medical Research Council. *Aids to the investigation of the peripheral nervous system*. (Her Majesty's Stationary Office, 1943).
- 3 Meucci, N., Cappellari, A., Barbieri, S. *et al.* Long term effect of intravenous immunoglobulins and oral cyclophosphamide in multifocal motor neuropathy. *J Neurol Neurosurg Psychiatry* **63**, 765-769 (1997).
- 4 Nobile-Orazio, E., Cappellari, A., Meucci, N. *et al.* Multifocal motor neuropathy: clinical and immunological features and response to IVIg in relation to the presence and degree of motor conduction block. *J Neurol Neurosurg Psychiatry* **72**, 761-766 (2002).
- 5 Van den Berg, L. H., Kerkhoff, H., Oey, P. L. *et al.* Treatment of multifocal motor neuropathy with high dose intravenous immunoglobulins: a double blind, placebo controlled study. *J Neurol Neurosurg Psychiatry* **59**, 248-252 (1995).
- 6 van Schaik, I. N., van den Berg, L. H., de Haan, R. *et al.* Intravenous immunoglobulin for multifocal motor neuropathy. *Cochrane database of systematic reviews*, CD004429 (2005).
- 7 Van den Berg-Vos, R. M., Van den Berg, L. H., Franssen, H. *et al.* Treatment of multifocal motor neuropathy with interferon-beta1A. *Neurology* **54**, 1518-1521 (2000).
- 8 van Swieten, J. C., Koudstaal, P. J., Visser, M. C. *et al.* Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* **19**, 604-607 (1988).
- 9 Azulay, J. P., Blin, O., Pouget, J. *et al.* Intravenous immunoglobulin treatment in patients with motor neuron syndromes associated with anti-GM1 antibodies: a double-blind, placebo-controlled study. *Neurology* **44**, 429-432 (1994).
- 10 Baumann, A., Hess, C. W., and Sturzenegger, M. IVIg dose increase in multifocal motor neuropathy: a prospective six month follow-up. *J Neurol* **256**, 608-614 (2009).
- 11 Cats, E. A., van der Pol, W. L., Piepers, S. *et al.* New liquid intravenous immunoglobulin (10 % IVIg) for treatment of multifocal motor neuropathy: a prospective study of efficacy, safety and tolerability. *J Neurol* **255**, 1598-1599 (2008).
- 12 Leger, J. M., Chassande, B., Musset, L. *et al.* Intravenous immunoglobulin therapy in multifocal motor neuropathy: a double-blind, placebo-controlled study. *Brain* **124**, 145-153 (2001).
- 13 van Nes, S. I., Faber, C. G., and Merkies, I. S. Outcome measures in immune-mediated neuropathies: the need to standardize their use and to understand the clinimetric essentials. *J Peripher Nerv Syst* **13**, 136-147 (2008).
- 14 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-59 (2006).
- 15 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* **49**, 711-717 (1996).
- 16 van Schaik, I. N., Bouche, P., Illa, I. *et al.* European Federation of Neurological Societies/Peripheral Nerve Society guideline on management of multifocal motor neuropathy. *Eur J Neurol* **13**, 802-808 (2006).
- 17 Cats, E. A., van der Pol, W. L., Piepers, S. *et al.* Correlates of outcome and response to IVIg in 88 patients with multifocal motor neuropathy. *Neurology* **75**, 818-825 (2010).
- 18 Bombardier, C. and Tugwell, P. A methodological framework to develop and select indices for clinical trials: statistical and judgmental approaches. *J Rheumatol* **9**, 753-757 (1982).
- 19 Streiner, D. L. and Norman, G. R. *Health measurement scales. A practical guide to their development and use*. (Oxford University Press, 2nd ed, 1998).
- 20 W.H.O. The International Classification of Functioning, Disability and Health (ICF). (WHO, Geneva, 2010).
- 21 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 22 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Clinimetric evaluation of a new overall disability scale in immune mediated polyneuropathies. *J Neurol Neurosurg Psychiatry* **72**, 596-601 (2002).
- 23 Andrich, D., Lyne, A., Sheridan, B. *et al.* *Rasch Unidimensional Measurement Models (RUMM2020 Version 4.0)*. (2003).
- 24 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1960).

- 25 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 26 Hermans, M. C., Merkies, I. S., Laberge, L. *et al.* Fatigue and daytime sleepiness scale in myotonic dystrophy type 1. *Muscle & nerve* **47**, 89-95 (2013).
- 27 Fisher, W. P. Reliability statistics. *Rasch Meas Transact* **6**, 238 (1992).
- 28 Linacre, J. M. Sample Size and Item Calibration Stability. *Rasch Measurement Transactions* **7** (1994).
- 29 Wright, B. D. Rack and Stack: Time 1 vs. Time 2. *Rasch Measurement Transactions* **17**, 905-906 (2003).
- 30 Wright, B. D. and Stone, M. H. *Best test design*. (Media Press, 1979).
- 31 Hobart, J. C., Cano, S. J., and Thompson, A. J. Effect sizes can be misleading: is it time to change the way we measure change? *J Neurol Neurosurg Psychiatry* **81**, 1044-1048 (2010).
- 32 Lai, J.-S. and Eton, D. T. Clinically meaningful gaps. *Rasch Meas Transact* **15**, 850 (2002).
- 33 Bland, J. M. and Altman, D. G. Multiple significance tests: the Bonferroni method. *Bmj* **310**, 170 (1995).
- 34 Federico, P., Zochodne, D. W., Hahn, A. F. *et al.* Multifocal motor neuropathy improved by IVIg: randomized, double-blind, placebo-controlled study. *Neurology* **55**, 1256-1262 (2000).
- 35 Beaton, D. E., Boers, M., and Wells, G. A. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current opinion in rheumatology* **14**, 109-114 (2002).
- 36 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health technology assessment* **13**, iii, ix-x, 1-177 (2009).
- 37 Vanhoutte, E. K., Faber, C. G., and Merkies, I. S. 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscular disorders : NMD* **23**, 924-933 (2013).
- 38 Umapathi, T., Hughes, R. A., Nobile-Orazio, E. *et al.* Immunosuppressant and immunomodulatory treatments for multifocal motor neuropathy. *Cochrane database of systematic reviews* **4**, CD003217 (2012).
- 39 Van den Berg, L. H., Franssen, H., and Wokke, J. H. The long-term effect of intravenous immunoglobulin treatment in multifocal motor neuropathy. *Brain : a journal of neurology* **121** ( Pt 3), 421-428 (1998).
- 40 Stevens, S. S. On the theory of scales of measurement. *Science* **103**, 677-680 (1946).
- 41 Merbitz, C., Morris, J., and Grip, J. C. Ordinal scales and foundations of misinference. *Archives of physical medicine and rehabilitation* **70**, 308-312 (1989).



# Changing outcome in inflammatory neuropathies: Rasch–comparative responsiveness

*Thomas H.P. Draak<sup>1</sup>, MD; Els K. Vanhoutte<sup>1</sup>, MD; Sonja I. van Nes<sup>2</sup>, MD, PhD; Kenneth C. Gorson<sup>3</sup>, MD;  
W. Ludo Van der Pol<sup>4</sup>, MD, PhD; Nicolette C. Notermans<sup>4</sup>, MD, PhD; Eduardo Nobile-Orazio<sup>5</sup>, MD, PhD;  
Jean-Marc Léger<sup>6</sup>, MD, PhD; Peter Y.K. Van den Bergh<sup>7</sup>, MD, PhD; Giuseppe Lauria<sup>8</sup>, MD, PhD;  
Vera Bril<sup>9</sup>, MD, PhD; Hans Katzberg<sup>9</sup>, MD; Michael P.T. Lunn<sup>10</sup>, MD, PhD; Jean Pouget<sup>11</sup>, MD, PhD;  
Anneke J. van der Kooi<sup>12</sup>, MD, PhD; Angelika F. Hahn<sup>13</sup>, MD, PhD; Pieter A. van Doorn<sup>2</sup>, MD, PhD;  
David R. Cornblath<sup>14</sup>, MD; Leonard H. van den Berg<sup>4</sup>, MD, PhD; Catharina G. Faber<sup>1</sup>, MD, PhD;  
Ingemar S.J. Merkies<sup>1,15</sup>, MD, PhD; on behalf of the PeriNomS study group*

<sup>1</sup>Department of Neurology, University Medical Centre Maastricht, Maastricht, the Netherlands

<sup>2</sup>Department of Neurology, Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands

<sup>3</sup>Department of Neurology, St. Elizabeth's Medical Center, Tufts University School of Medicine, Boston, USA; <sup>4</sup>Department of Neurology, Rudolf Magnus Institute of Neuroscience University

Medical Centre Utrecht, the Netherlands; <sup>5</sup>Department of Neurological Sciences, Milan; University, Humanitas Clinical Institute, Rozzano, Milan, Italy; <sup>6</sup>Department of Neurology, Hôpital

de la Salpêtrière, Paris, France; <sup>7</sup>Department of Neurology, Catholique University of Louvain,

Belgium; <sup>8</sup>Department of Clinical Neurosciences, 3<sup>rd</sup> Neurology Unit, Milan, Italy; <sup>9</sup>Department of Neurology, Toronto General Hospital, Toronto, Canada; <sup>10</sup>Department of Neurology, Centre

for Neuromuscular disease, National Hospital for Neurology and Neurosurgery, Queen Square, London, United Kingdom; <sup>11</sup>Department of Neurology, Centre de Référence des Maladies

Neuromusculaires et de la SLA Hôpital de La Timone, Marseille, France; <sup>12</sup>Department of Neurology, Academic Medical Centre, Amsterdam, The Netherlands; <sup>13</sup>Department of Neurology,

London Health Science Center, London, Ontario, Canada; <sup>14</sup>Department of Neurology, Johns Hopkins School of Medicine, Baltimore, USA; <sup>15</sup>Department of Neurology, Spaarne Hospital,

Hoofddorp, the Netherlands.

## Abstract

We performed responsiveness comparison between the patient-reported inflammatory-Rasch-built Overall Disability Scale (I-RODS) and the widely used clinician-reported Inflammatory Neuropathy Cause and Treatment-Overall Neuropathy Limitation Scale (INCAT-ONLS) in patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and immunoglobulin M-monoclonal gammopathy of undetermined significance related polyneuropathy (IgM-MGUSP).

One hundred thirty-seven patients (GBS: 55, CIDP: 59, IgM-MGUSP: 23) with a new diagnosis or clinical relapse assessed both scales. Patients with GBS/CIDP were examined at 0, 1, 3, 6, and 12 months; patients with IgM-MGUSP at 0, 3, and 12. We subjected all data to Rasch analyses, and calculated for each patient the magnitude of change on both scales using the minimum clinically important difference (MCID) related to the individual standard errors (SE). A responder was defined as having a  $MCID-SE \geq 1.96$ . Individual scores on both measures were correlated with the EuroQoL-thermometer (heuristic responsiveness).

The I-RODS showed a significantly higher proportion of meaningful improvement compared with the INCAT-ONLS findings in GBS/CIDP. For IgM-MGUSP, the lack of responsiveness during the 1-year study did not allow a clear separation. Heuristic responsiveness was consistently higher with the I-RODS.

The I-RODS more often captures clinically meaningful changes over time, with a greater magnitude of change, compared to the INCAT-ONLS disability scale in patients with GBS and CIDP. The I-RODS offers promise for being a more sensitive measure and its use is therefore suggested in future trials involving patients with GBS and CIDP.

## Introduction

Most randomized controlled trials in patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and monoclonal gammopathy of undetermined significance related polyneuropathy (MGUSP) have used primary outcome measures that are ordinal-based, despite their known limitations.<sup>1-8</sup> One of the most widely used scales in these diseases is the 10-point Inflammatory Neuropathy Cause and Treatment (INCAT) Disability Scale,<sup>9</sup> and its latest modified 12-point Overall Neuropathy Limitation Scale (INCAT-ONLS).<sup>10</sup> Both scales are clinician reported outcome measures at ordinal level. To overcome the shortcomings of ordinal-based outcome measures, we previously developed a 24-item Rasch-built Overall Disability Interval Scale (RODS). This patient reported outcome measure captures activity limitations and participation restrictions at the moment of completion and was specifically designed for patients with GBS, CIDP, and MGUSP.<sup>11</sup> In this report, we refer to the RODS as the Inflammatory RODS (I-RODS). Despite having high internal/external validity, proper reliability scores, and high discriminatory validity, the responsiveness of the I-RODS has not been examined.<sup>11</sup> We investigated the ability of the I-RODS to capture clinically meaningful changes over time in serially examined newly diagnosed or relapsing patients with GBS, CIDP, and immunoglobulin M (IgM)-MGUSP. In addition, we compared the responsiveness between the I-RODS and the INCAT-ONLS scale to determine which scale is superior. Finally, we performed cross-cultural evaluation for the most responsive scale.

## Methods

### Patients

This study is part of the Peripheral Neuropathy Outcome Measures Standardization (PeriNomS) Study (duration: 7 years; database closed December 31, 2012), an international collaborative effort of 26 neuromuscular centres with special interest in inflammatory neuropathies, located in 9 countries (USA, Canada, Spain, Brazil, Italy, France, Belgium, the Netherlands, and UK). The primary aim of the PeriNomS study is to select the most responsive outcome measures in GBS/CIDP/IgM-MGUSP and to develop a core set of outcome measures for each illness to be used in future clinical studies.

The participating centres recruited and serially examined a total of 137 patients with newly diagnosed (or clinically relapsing) GBS, CIDP, and IgM-MGUSP during a 12-month follow-up period. Patients were treated at the discretion of their physician.

### Eligibility, Standard Protocol Approvals, Registrations, and Patient Consents

Eligibility included normal cognitive function, age 18 years or older, fulfilling international criteria for the diagnosis GBS, CIDP, or IgM (preferably anti-MAG+)-

MGUSP,<sup>12-14</sup> having clinical deterioration as judged by the patient and physician, and not receiving immune-therapy for their polyneuropathy at least 2 months prior to inclusion. Patients were excluded if other causes of polyneuropathy (e.g. renal insufficiency, diabetes mellitus) or any other conditions (e.g. alcohol abuse: >5 IU/day) that might interfere with nervous system and physical functioning (e.g. arthritis) were present.

All patients gave written informed consent before participation in the study. The local medical ethics committee in each participating center approved the protocol.

### **Rasch method and Outcome assessments**

The Rasch model is a modern technique that enables the transformation of ordinal-based outcome measures to the interval level.<sup>11, 15, 16</sup> Rasch states that the probability that a person may affirm an item/task depends on the item difficulty and person's ability level (degree of illness). Readers are invited to a more educational explanation of the Rasch methodology using the supplementary material of the original I-RODS report.<sup>11, 16</sup> The I-RODS comprises 24-items which address daily and social activities (easiest item: ability to read a book/newspaper; most difficult item: ability to run) and has 3 response options: (0) impossible to perform, (1) performed with difficulty, and (2) easily performed.<sup>11</sup> The I-RODS is easy to complete (in 2-3 minutes) by the patient or caregiver and the scores are at the interval level after Rasch-method transformation, thus having a high degree of data recruitment precision.<sup>17</sup> Its scores are expressed in log-odds (logits).<sup>11, 16</sup>

The INCAT-ONLS, has never been clinimetrically evaluated, but 2 very similar scales have demonstrated their scientific soundness that can be extrapolated to the INCAT scale.<sup>10, 18</sup> The INCAT-ONLS provides good functional descriptions in a checklist format, suitable for interviewing patients. Daily activities such as dressing the upper part of the body, (un)doing buttons/zippers, washing/brushing hair, using a knife/fork, turning a key in a lock, and problems with walking are assessed, and the results are summated into an arm grade (0-5) and leg grade (0-7) (total INCAT-ONLS score ranges from 0 [no signs of disability] to 12 [most severe disability]).<sup>9, 10</sup>

### **Procedures and Statistics**

- **General aspects**

The I-RODS and INCAT-ONLS were assessed at 0/1/3/6/12 months in patients with GBS/CIDP; because of the more indolent disease course, patients with IgM-MGUSP were examined at 0/3/12 months.<sup>14</sup>

We provided patients and physicians with standardized instructions how to complete the outcome measures. The data from both scales, including the following person factors, were subjected to Rasch analyses: gender, age (<50 vs 50-65 vs >65 years; cutoffs chosen aiming for approximately equal distribution), and illness (GBS vs CIDP vs IgM-MGUSP). It was not possible to assess each country as an individual

person factor because the number of patients recruited per country was generally low, except for the Netherlands (61% were Dutch). Therefore, the following cultural regions were defined and entered: living in the Netherlands (yes/no), Europe (yes/no), with further geographic divisions that included northern Europe (United Kingdom, Belgium, the Netherlands) vs southern Europe (Italy, France, Spain), vs North and South America (United States, Canada, Brazil).

- **Responsiveness comparison strategies:**

We have chosen 2 strategies to compare responsiveness for both scales in the serially examined patients:

- Strategy 1. We performed a responsiveness comparison at the individual person level using individual standard errors (SEs) after subjecting the data to the Rasch method, thereby constructing interval measures from counts of qualitatively-ordered observations.<sup>15, 16, 19</sup> SEs may vary across the theoretical range of an outcome measure, and therefore the significance of a clinically meaningful change in any particular patient may also vary.<sup>15, 20-22</sup> The minimally clinically important difference (MCID)-related SE (MCID-SE) score is based on the previously described significant changes.<sup>20</sup> Briefly, MCID-SE was computed for each Rasch-transformed scale (I-RODS vs INCAT-ONLS) in each of the 137 examined patients separately: (1) their personal change (personal-location at *i* month minus personal location at baseline [entry]; where *i* = 1, 3, 6 or 12 months of follow-up in GBS/CIDP, and 3 and 12 months in IgM-MGUSP); (2) the corresponding SE of difference related to their individual change ( $SE_{diff} = \text{square-root} [SE_{entry2} + SE_{i\text{ month}2}]$ ); and (3) the final MCID-SE calculations were computed by dividing the individual change scores by their corresponding  $SE_{diff}$  ( $MCID-SE = [\text{personal-location at } i \text{ month minus personal-location at entry}] / SE_{diff}$ ).<sup>20</sup> Subsequently, the calculated MCID-SE findings were divided into the following subgroups: subgroup 1 (clinically important improvement):  $MCID-SE \geq 1.96$ ; subgroup 2 (clinically unimportant improvement):  $0 < MCID-SE < 1.96$ ; subgroup 3 (no change):  $MCID-SE = 0$ ; subgroup 4 (clinically unimportant deterioration):  $-1.96 < MCID-SE < 0$ ; and subgroup 5 (clinically important deterioration):  $MCID-SE \leq -1.96$ ).<sup>20</sup> Figures e-1 and e-2 on the Neurology® Web site at Neurology.org explain the use of the individually obtained SE through Rasch to determine “being a responder”.
- Strategy 2. We took a more patient-oriented approach to determine the so-called “heuristic responsiveness” for both scales.<sup>23, 24</sup> Heuristic techniques are based upon comparing changes using a scale with an external indicator completed by patients.<sup>23, 24</sup> In the current study, we have chosen the personal patient-derived EuroQoL “thermometer”, since this ruler helps patients to indicate their personal health status (range: 0 [worst state] to 100 [best state]).<sup>25</sup> These scores were correlated with the scores (patient’s ability, [logits]) of each individual patient after subjecting the data for both scales to Rasch.



- **Statistics for strategy 1.**

The proportion of patients reaching clinically important improvement on both scales was compared in GBS, CIDP, and IgM-MGUSP separately. An accepted cutoff value for the definition of a clinically important improvement is  $\geq 1.96 \times$  personal obtained SE, at a chosen specificity of 95% confidence interval.<sup>15, 20, 22</sup> Subsequently, as previously described, we created Kaplan–Meier curves to estimate and compare the cumulative proportion of patients per illness (overall, and separately in GBS, CIDP, and IgM-MGUSP) demonstrating significant improvement over time (at 1, 3, 6, and 12 months of follow-up) for both scales (I-RODS vs INCAT-ONLS).<sup>26</sup> The log-rank test was used to examine possible differences.

- **Statistics for Strategy 2.**

Univariate regression studies using splines (to improve data fitting) were performed to determine a possible correlation between the Rasch-transformed personal scores and the EuroQoL thermometer for both scales (correlation presented as  $R^2$ : the fraction of variance explained by the independent variables).<sup>27</sup>

- **Cross-cultural evaluation of the most responsive scale**

Immune-mediated neuropathies are rare disorders and therefore most randomized trials in these patients have been completed through multicenter international studies. A prerequisite is that outcome measures used internationally in any condition must overcome general cross-cultural differences.<sup>28, 29</sup> Thus far, no outcome measure in inflammatory neuropathy studies has been cross-culturally validated. In the current study, the person factors ‘the Netherlands’, ‘Europe’, and ‘culture-difference’ were used to determine the cultural validity of the most responsiveness scale (I-RODS or INCAT-ONLS).<sup>28</sup> In essence, data from the various regions is compared through differential item functioning (DIF) analyses using the 95th confidence interval boundaries to determine invariance as has been addressed previously as an acceptable surrogate marker to examine cultural differences.<sup>28</sup> Invariance suggests that the probability of a patient in one region completing an item will be the same as the probability of a patient in another region affirming the same item, given that they are both at the same trait level (having the same degree of functional deficit).<sup>28, 30</sup>

- **Software**

We performed Rasch analyses using RUMM2030 with a partial credit model as default.<sup>31</sup> The I-RODS data were easily subjected to RUMM2030, since all 24 items are straightforward with equal (3) response options. For the INCAT-ONLS, we took 2 approaches:

1. The arm grade and leg grade scores were subjected to RUMM2030 as if the INCAT-ONLS scale was composed by these 2 independent items; this approach was chosen as the arm and leg grades have been used to detect meaningful changes after an intervention;
2. The INCAT-ONLS was also redefined to its original 5 arm activities plus 2 items addressing mobility (1 - do you have walking difficulty: no/yes; and 2 - how do

you usually mobilize outdoors: independently/with 1 stick/with 2 sticks/with a wheelchair). These 7 items were subjected to RUMM2030 as a group of individual items.

Hence, the I-RODS Rasch data were compared with both Rasch-transformed (RT-2-item and RT-7-item) INCAT-ONLS approaches. Further analyses were performed using Stata for Windows XP (version 12.0, StataCorp, Texas, USA). Differences in possible ceiling and floor percentages between scales were also examined (X<sup>2</sup>-test; p-value <0.05 was accepted as statistically significant).

**Table 1.** General characteristics of longitudinal recruited patients (n=137)

Diagnosis	GBS	CIDP	IgM-MGUSP
Number of patients	55	59	23
<b>Age (years)</b>			
mean (SD)	53.4 (17.7)	52.7 (14.7)	62.8 (10.6)
range	19-90	18-79	45-85
<b>Gender, number (%)</b>			
Female	18 (32.7)	17 (28.8)	8 (34.8)
Male	37 (67.3)	42 (71.2)	15 (65.2)
<b>Country of assessment</b>			
the Netherlands	25	20	8
USA	6	11	4
Italy	9	9	3
Spain	-	2	-
France	1	4	1
Belgium	2	2	5
United Kingdom	2	2	1
Canada	8	7	1
Brazil	2	2	-
<b>Current medication (%)*</b>			
IVIg and/or Steroids	81.8%	67.8%	34.8%
Plasma Exchange	16.4%	5.1%	17.4%
Immunosuppressive	-	3.4%	13%
Other	-	10.2%	34.8%

**Legend to Table 1:** CIDP= chronic inflammatory demyelinating polyradiculoneuropathy; GBS = Guillain-Barré syndrome; IgM-MGUSP= IgM monoclonal gammopathy of undetermined significant related polyneuropathy; IVIg = intravenous immunoglobulin. \*Some missing data on medication use.

## Results

### Patients

The demographics of the 137 patients are presented in Table 1. Nineteen of the 23 patients with IgM-MGUSP had anti-MAG antibodies. There were a total of 600 patient visits, and 94% of the patients completed 1-year follow-up.

## Comparison I-RODS versus INCAT-ONLS

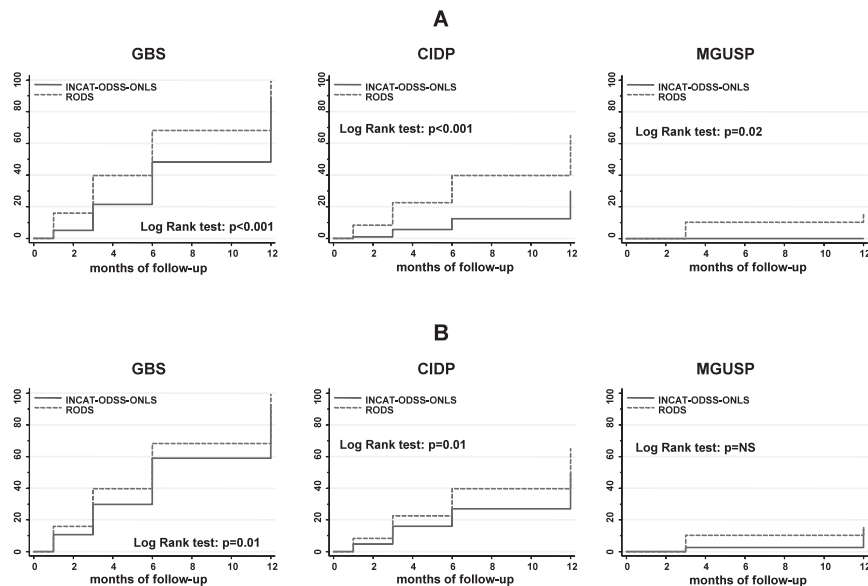
- General aspects:

The I-RODS had overall lower SEs than seen with the INCAT-ONLS (Figure e-3) and also demonstrated significantly lower ceiling scores (I-RODS ceiling: 9.5% of all records, Rasch transformed-2-item INCAT-ONLS: 17.4%, RT-7-item INCAT-ONLS: 20.8%;  $\chi^2$ -tests:  $p < 0.001$ ). Both measures demonstrated minimum floor effects, but these were not significantly different (I-RODS: 2.9%; 2-item INCAT-ONLS: 1.7%; 7-item INCAT-ONLS: 2.9%).

## Comparison at the individual personal level using MCID-SE cut-off:

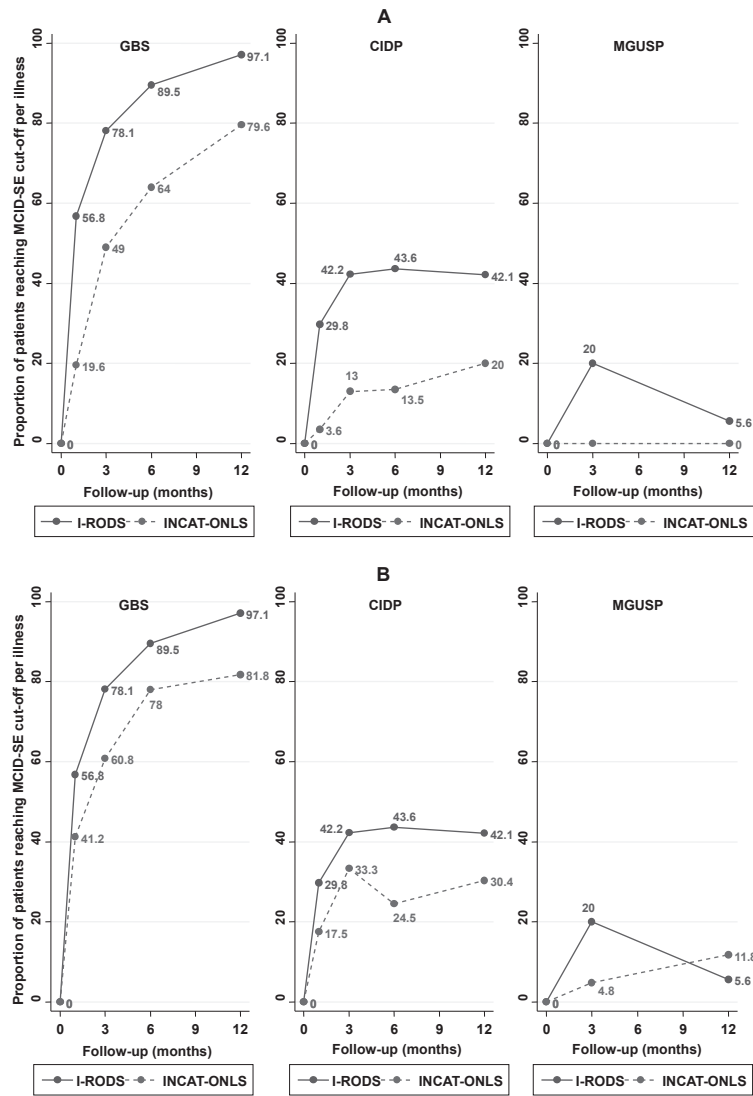
A significantly higher cumulative proportion of patients with GBS/CIDP demonstrated a meaningful improvement (MCID-SE cut-off  $\geq 1.96$ ) when assessed with the I-RODS versus the INCAT-ONLS (figure 1;  $p$ -values:  $p = 0.01$  to  $p < 0.001$ , favoring the I-RODS).

**Figure 1.** Comparing improvement between the I-RODS vs (A) the RT-2-item INCAT-ONLS and (B) the RT-7-item INCAT-ONLS



**Legend to Figure 1:** A consistently higher proportion of patients reached a significant meaningful improvement (MCID-SE cutoff  $\geq 1.96$ ) when assessed with the I-RODS vs the INCAT-ONLS. In patients with IgM-MGUSP, the number of patients with a significant change was very low. CIDP = chronic inflammatory demyelinating polyradiculoneuropathy; GBS = Guillain-Barré syndrome; IgM-MGUSP = immunoglobulin M-monoclonal gammopathy of undetermined significance related polyneuropathy; INCAT = Inflammatory Neuropathy Cause and Treatment; I-RODS = Inflammatory Rasch-built Overall Disability Scale; MCID-SE = 5 minimal clinically important difference–standard error; NS = not significant; ONLS = Overall Neuropathy Limitation Scale; RT = Rasch-transformed; RT-2 = Raschtransformed; INCAT-ONLS composed by 2 items = the arm grade with 5 response options and the leg grade with 7 response options; RT-7 = Raschtransformed INCAT-ONLS composed by 7 items: the original 5 arm activities plus 2 items addressing mobility (1—do you have walking difficulty: no/yes; and 2—how do you usually mobilize outdoors: independently/with 1 stick/with 2 sticks/with a wheelchair).

**Figure 2.** Comparison in improvement between the I-RODS vs the (A) RT-2-item INCAT-ONLS and (B) RT-7-item INCAT-ONLS



**Legend to Figure 2:** RT-2: Rasch-transformed INCAT-ONLS composed by 2 items: the arm grade with 5 response options and the leg grade with 7 response options. RT-7: Rasch-transformed INCAT-ONLS composed by 7 items: the original 5 arm activities plus 2 items addressing mobility (1—do you have walking difficulty: no/yes; and 2—how do you usually mobilize outdoors: independently/with 1 stick/with 2 sticks/with a wheelchair). CIDP = chronic inflammatory demyelinating polyradiculoneuropathy; GBS = Guillain-Barré syndrome; INCAT = Inflammatory Neuropathy Cause and Treatment; I-RODS 5 Inflammatory Rasch-built Overall Disability Scale; MCID-SE = minimal clinically important difference–standard error; MGUSP = monoclonal gammopathy of undetermined significance-related polyneuropathy; ONLS = Overall Neuropathy Limitation Scale; RT = Rasch-transformed.

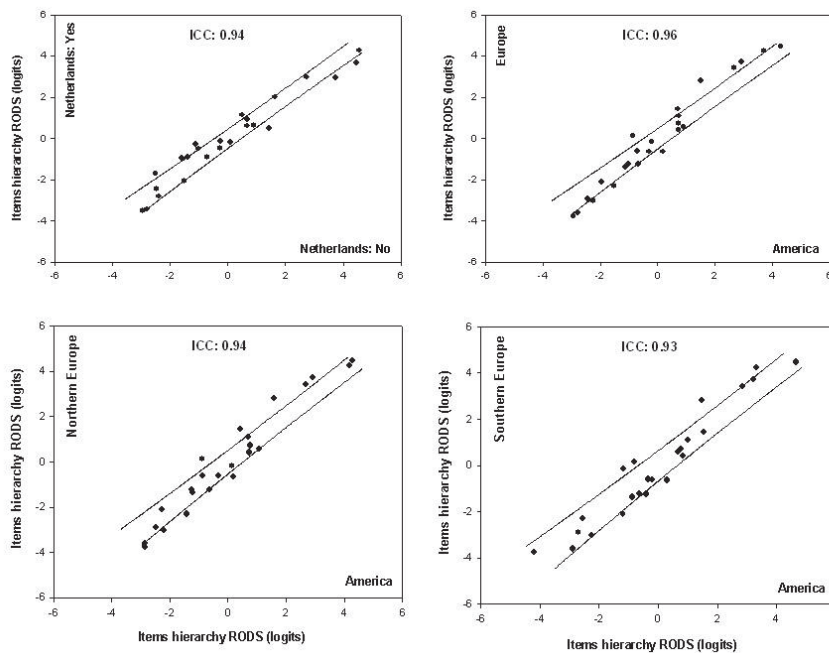
In IgM-MGUSP, only a small proportion of these patients reached this cut-off (<20%) in both scales (figure 1;  $p=0.02$  favoring the I-RODS vs RT-2-item INCAT-ONLS). The responsiveness was consistently higher in GBS when compared to the other illnesses. At 1-year follow-up, 97% of patients with GBS reached the MCID-SE cut-off scores using the I-RODS vs 80% for the RT-2-item INCAT-ONLS and 82% for the RT-7-item INCAT-ONLS; in CIDP these numbers were 42% (I-RODS) vs 20% (RT-2-item INCAT-ONLS) and 30% (RT-7-item INCAT-ONLS) (figure 2).

Figure e-4 presents an overall view of the magnitude of changes from another perspective seen of the I-RODS versus the INCAT-ONLS, being consistently higher for the I-RODS.

### Comparison through heuristic responsiveness

The correlation found between the EuroQol-thermometer and patient's ability is presented in Table 2. The association between the I-RODS and EuroQol-thermometer was consistently higher when compared with the INCAT-ONLS-EuroQol-thermometer associations. The obtained values were higher in patients with GBS>CIDP>IgM-MGUSP.

**Figure 3.** The invariance of the items' hierarchy of the I-RODS across countries and regions



**Legend to Figure 3:** Most items of the I-RODS fell within the 95%-confidence interval (solid) lines indicating that most items are ranked with the same hierarchy between the groups being compared and show no DIF; easiest items are at the bottom left part and most difficult items at the right top part. However, some items fall outside the 95%-CI. A total of 5 of the 24 items showed DIF. DIF: differential item functioning; ICC: intra-class correlation coefficient; I-RODS: inflammatory Rasch-built overall disability scale.

**Table 2.** Correlations between EuroQoL-thermometer and the I-RODS and INCAT-ONLS outcome measures

	GBS				CIDP				IgM-MGUSP			
	Overall		2-item		7-item		2-item		7-item		2-item	
	I-RODS	0.52	RT-INCAT-ONLS	0.45	RT-INCAT-ONLS	0.44	I-RODS	0.61	RT-INCAT-ONLS	0.54	RT-INCAT-ONLS	0.30
<b>EuroQoL thermometer</b>												
<b>Corresponding p-value</b>		$p<0.0001$		$p<0.0001$		$p<0.0001$		$p<0.0001$		$p<0.0001$		$p<0.0001$

**Legend to Table 2:** Abbreviations: CIDP = chronic inflammatory demyelinating polyradiculoneuropathy; GBS = Guillain-Barré syndrome; IgM-MGUSP = immunoglobulin M-monoclonal gammopathy of undetermined significance related polyneuropathy; INCAT-ONLS = Inflammatory Neuropathy Cause and Treatment–Overall Neuropathy Limitation Scale; I-RODS = Inflammatory Rasch-built Overall Disability Scale; RT = Rasch-transformed.

The EuroQoL thermometer captures the quality of life feeling of the individual patient at that moment, and relating the activity and participation scale to this ruler provides a correlation with the individual patient's own judgment of their clinical state. The strength of association is presented as R2: the fraction of variance explained by the independent variables from a regression model.

### Cross-cultural validation

Subsequently, we examined the cross-cultural validity of the I-RODS as the most responsiveness scale (figure 3). Most items were within the 95%-confidence interval boundaries indicating proper hierarchical ordering and had no DIF. However, 5 items showed DIF on several cross-cultural factors: one item ("able to travel by public transport") showed uniform-DIF on person factor "the Netherlands"; two items ("able to walk outdoors" and "able to move a chair") had uniform-DIF on "Europe"; and two items ("able to wash lower body" and "able to walk and avoid obstacles") had uniform-DIF on "cultural-difference".

### Discussion

We found that the I-RODS was a more responsive scale in GBS and CIDP compared to the widely used ordinal-based INCAT-ONLS using the MCID concept related to the varying standard errors seen throughout the Rasch-transformed scales' continuum. Based on the results in the current study, experts in the field recently suggested the I-RODS as the primary outcome measure for future trials involving patients with GBS and CIDP.<sup>21</sup>

In this study <20% of IgM-MGUSP patients demonstrated clinical changes. The number of IgM-MGUSP patients in this study was small, the medical interventions varied, and few patients had any clinically meaningful response to any treatment. These factors have made it exceptionally difficult to assess responsiveness in this indolent disease. More importantly, our data suggest that 1-year follow-up may be insufficient to assess clinically meaningful changes in this disorder. This observation is especially relevant because prior treatment trials in IgM-MGUSP have had a short duration (< 1 year).<sup>4, 32</sup>

The INCAT-ONLS was transformed through Rasch into a 2-item and 7-item scale, creating a higher level of measurement precision for the INCAT-ONLS data and thereby bypassing the shortcomings of ordinal-based scales.<sup>7, 8</sup> The INCAT-ONLS remained less responsive compared to the I-RODS. The findings in the current study may also be seen as a new way to define "being a responder" for future trial design in inflammatory neuropathies and may serve as a frame for other neurological illnesses, facing the same ordinal-based deficiencies,<sup>20, 33</sup> as well as for other disciplines (e.g., rheumatology) treating patients with chronic conditions.

The results of this and other studies<sup>17, 34, 35</sup> pose a considerable and much broader question: Why do neurologists continue to use ordinal measures in clinical trials despite the known constraints? Several facets may explain the improper use of ordinal measures: first, most neurologists consider observations in clinical setting as real measurements and are unaware that quantitative observations are based on counting observed events, while meaningful measurements are based on the arithmetical properties of interval scales or ratio scales.<sup>7, 17, 36</sup> Observations, such as daily activities as part of the INCAT-ONLS with their Likert-type response options, are thus ordinal

and provide a good description of the clinical question of interest. The problem arises when we start giving the ordinal descriptions numbers arbitrarily, create sum scores, treating these scores as linear, and make assumptions from the findings.<sup>7, 37, 38</sup> Second, the current medical curriculum does not provide training in clinimetrics, the science that addresses measurement requirements in the clinical setting, and therefore, the exposure of this knowledge to physicians is relatively poor. As a consequence, experts in any medical field or illness generally provide information regarding “the best way” to assess clinical changes in a particular illness of interest, assuming, often incorrectly, that they have the proper knowledge to do so. Subsequently, the risk of drawing incorrect inferences from trials with ordinal measures as the primary outcome becomes high, but this remains generally unacknowledged.<sup>17, 34</sup> Positive trials may be falsely positive and vice versa. Comparison of raw I-RODS (with an arbitrarily taken cut-off) versus ordinal INCAT-ONLS (using a 1-point cut-off as often suggested) is therefore considered a less proper solution. We simply do not know what we are measuring and the comparison becomes uncertain.<sup>17, 38, 39</sup> Third, when being exposed to modern methods like the Rasch and Item Response Theory, physicians generally may become somewhat uncomfortable and opposed to these techniques mostly due to their less accessible arithmetical properties. This may contribute to a “holding-on attitude” of physicians to the seemingly more intuitive, but less proper ordinal-based measures. Reporting results in future trials could shift from comparing two groups through *p* values to reporting proportions of patients demonstrating individual meaningful improvement through MCID-SE method.

There are limitations to our study that should be addressed. First, it would have been ideal to have at least 50 patients per illness and country included in the study to properly evaluate the cross-cultural validation of the I-RODS. Because GBS, CIDP, and IgM-MGUSP are rare disorders, recruitment was challenging. Second, although the I-RODS scale is easily applicable, some caution is warranted as 5 of its 24 items demonstrated item bias in different geographical regions and the applicability of these items to different settings needs further cross-cultural exploration. Third, although responsiveness was demonstrated using the MCID-SE technique and the heuristic approach using the EuroQol-VAS, further anchor-based evaluation may be desirable to extent the comparison with existing instruments. These points are part of the future scope of the PeriNomS Study Group. Fourth, efforts are under way to validate the I-RODS by again presenting its content to focussed groups of GBS/CIDP patients and to examine the various ways the I-RODS could be used in future studies. Patients may be encouraged to indicate which target (level of functionality) they want to reach on the I-RODS ruler, the so-called patient’s Functional Acceptable Clinical Threshold (FACT) that may vary with their changing clinical condition in time.<sup>40</sup> Despite these limitations, the I-RODS is the only outcome measure at the interval level that has demonstrated acceptable cultural validation in inflammatory neuropathies, in addition to having higher responsiveness scores when compared to the INCAT-ONLS.



The use of the I-RODS offers promise for being a more sensitive measure in future trials involving patients with GBS and CIDP. Additional studies in IgM-MGUSP are needed, as the responsiveness in this indolent condition was relatively poor.

## References

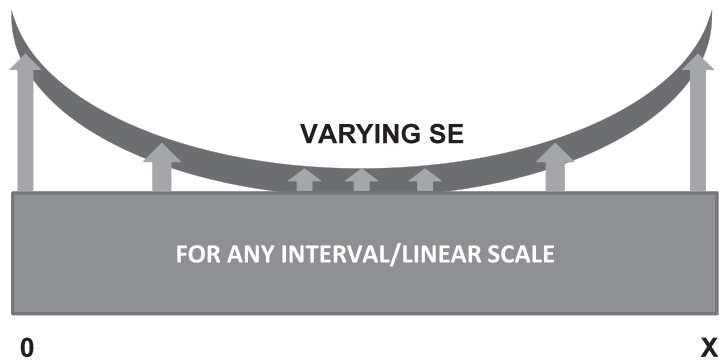
1. Hughes RA, Donofrio P, Bril V, et al. Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet neurology* 2008;7:136-144.
2. Dalakas MC, Rakocevic G, Salajegheh M, et al. Placebo-controlled trial of rituximab in IgM anti-myelin-associated glycoprotein antibody demyelinating neuropathy. *Annals of neurology* 2009;65:286-293.
3. van Koningsveld R, Schmitz PJ, Meche FG, Visser LH, Meulstee J, van Doorn PA. Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barre syndrome: randomised trial. *Lancet* 2004;363:192-196.
4. Leger JM, Viala K, Nicolas G, et al. Placebo-controlled trial of rituximab in IgM anti-myelin-associated glycoprotein neuropathy. *Neurology* 2013;80:2217-2225.
5. Nobile-Orazio E, Cocito D, Jann S, et al. Intravenous immunoglobulin versus intravenous methylprednisolone for chronic inflammatory demyelinating polyradiculoneuropathy: a randomised controlled trial. *Lancet neurology* 2012;11:493-502.
6. van Schaik IN, Eftimov F, van Doorn PA, et al. Pulsed high-dose dexamethasone versus standard prednisolone treatment for chronic inflammatory demyelinating polyradiculoneuropathy (PREDICT study): a double-blind, randomised, controlled trial. *Lancet neurology* 2010;9:245-253.
7. DeVellis RF. Classical test theory. *Med Care* 2006;44:S50-59.
8. Merkies IS, Lauria G, Faber CG. Outcome measures in peripheral neuropathies: requirements through statements. *Current opinion in neurology* 2012;25:556-563.
9. Hughes R, Bensa S, Willison H, et al. Randomized controlled trial of intravenous immunoglobulin versus oral prednisolone in chronic inflammatory demyelinating polyradiculoneuropathy. *Annals of neurology* 2001;50:195-201.
10. Graham RC, Hughes RA. A modified peripheral neuropathy scale: the Overall Neuropathy Limitations Scale. *J Neurol Neurosurg Psychiatry* 2006;77:973-976.
11. van Nes SI, Vanhoutte EK, van Doorn PA, et al. Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* 2011;76:337-345.
12. Van der Meche FG, Van Doorn PA, Meulstee J, Jennekens FG. Diagnostic and classification criteria for the Guillain-Barre syndrome. *Eur Neurol* 2001;45:133-139.
13. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of chronic inflammatory demyelinating polyradiculoneuropathy: report of a joint task force of the European Federation of Neurological Societies and the Peripheral Nerve Society—First Revision. *J Peripher Nerv Syst* 2010;15:1-9.
14. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of paraneoplastic demyelinating neuropathies. Report of a Joint Task Force of the European Federation of Neurological Societies and the Peripheral Nerve Society—first revision. *J Peripher Nerv Syst* 2010;15:185-195.
15. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1960.
16. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and rheumatism* 2007;57:1358-1362.
17. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation* 1989;70:857-860.
18. Merkies IS, Schmitz PJ, van der Meche FG, Samijn JP, van Doorn PA. Clinimetric evaluation of a new overall disability scale in immune mediated polyneuropathies. *J Neurol Neurosurg Psychiatry* 2002;72:596-601.
19. Hermans MC, Faber CG, De Baets MH, de Die-Smulders CE, Merkies IS. Rasch-built myotonic dystrophy type 1 activity and participation scale (DM1-Activ). *Neuromuscul Disord* 2010;20:310-318.
20. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health technology assessment (Winchester, England)* 2009;13:iii, ix-x, 1-177.
21. Vanhoutte EK, Faber CG, Merkies IS. 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscul Disord* 2013.
22. Heesch KC, Masse LC, Dunn AL. Using Rasch modeling to re-evaluate three scales related to physical activity: enjoyment, perceived benefits and perceived barriers. *Health Educ Res* 2006;21 Suppl 1:i58-72.

23. Liang MH. Evaluating measurement responsiveness. *J Rheumatol* 1995;22:1191-1192.
24. van Nes SI, Faber CG, Merkies IS. Outcome measures in immune-mediated neuropathies: the need to standardize their use and to understand the clinimetric essentials. *J Peripher Nerv Syst* 2008;13:136-147.
25. EuroQol—a new facility for the measurement of health-related quality of life. The EuroQol Group. *Health Policy* 1990;16:199-208.
26. van der Beek NA, Hagemans ML, van der Ploeg AT, van Doorn PA, Merkies IS. The Rasch-built Pompe-specific activity (R-PAct) scale. *Neuromuscul Disord* 2013;23:256-264.
27. Herndon JE, II, Harrell FE, Jr. The restricted cubic spline hazard model. *Comm Stat Theory Meth* 1990;19:639-663.
28. Kucukdeveci AA, Sahin H, Ataman S, Griffiths B, Tennant A. Issues in cross-cultural validity: example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. *Arthritis and rheumatism* 2004;51:14-19.
29. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *Journal of clinical epidemiology* 1993;46:1417-1432.
30. Andrich D. Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement* 1978;2:449-460.
31. Andrich D, Sheridan B, Luo G. Rasch models for measurement: RUMM2030. Perth 2010.
32. Lunn MP, Nobile-Orazio E. Immunotherapy for IgM anti-myelin-associated glycoprotein paraprotein-associated peripheral neuropathies. The Cochrane database of systematic reviews 2012;5:CD002827.
33. Knutsson I, Rydstrom H, Reimer J, Nyberg P, Hagell P. Interpretation of response categories in patient-reported rating scales: a controlled study among people with Parkinson's disease. *Health and quality of life outcomes* 2010;8:61.
34. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of clinical epidemiology* 1996;49:711-717.
35. Vanhoutte EK, Faber CG, van Nes SI, et al. Modifying the Medical Research Council grading system through Rasch analyses. *Brain : a journal of neurology* 2012;135:1639-1649.
36. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet neurology* 2007;6:1094-1105.
37. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? *J Rehabil Med* 2012;44:97-98.
38. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. *Archives of physical medicine and rehabilitation* 1989;70:308-312.
39. Stevens SS. On the theory of scales of measurement. *Science* 1946;103:677-680.
40. Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. *Arthrit Rheum-Arthr* 2006;55:526-530.31 De Land, P. N. and Chase, W. W. Entry I.J, types of data: nominal, ordinal, interval, and ratio scales. *Optometry and vision science : official publication of the American Academy of Optometry* 67, 155-156 (1990).
32. Grimby, G., Tennant, A., and Tesio, L. The use of raw scores from ordinal scales: time to end malpractice? *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* 44, 97-98 (2012).
33. Hobart, J. C., Cano, S. J., Zajicek, J. P. et al. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol* 6, 1094-1105 (2007).
34. Merbitz, C., Morris, J., and Grip, J. C. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil* 70, 308-312 (1989).
35. Stevens, S. S. On the theory of scales of measurement. *Science* 103, 677-680 (1946).
36. Knutsson, I., Rydstrom, H., Reimer, J. et al. Interpretation of response categories in patient-reported rating scales: a controlled study among people with Parkinson's disease. *Health Qual Life Outcomes* 8, 61 (2010).
37. Stucki, G., Daltroy, L., Katz, J. N. et al. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* 49, 711-717 (1996).
38. Vanhoutte, E. K., Faber, C. G., van Nes, S. I. et al. Modifying the Medical Research Council grading system through Rasch analyses. *Brain* 135, 1639-1649 (2012).

**Figure e-1.** Defining a responder using the varying individually obtained SE through Rasch analyses

**A      Defining a responder through Rasch analyses**

Depending on varying SE



**B                                      Method**

- Comparison between the scales was performed after transformation in RUMM2030
- The obtained locations with corresponding personal Standard Errors (SE's) were used for comparison

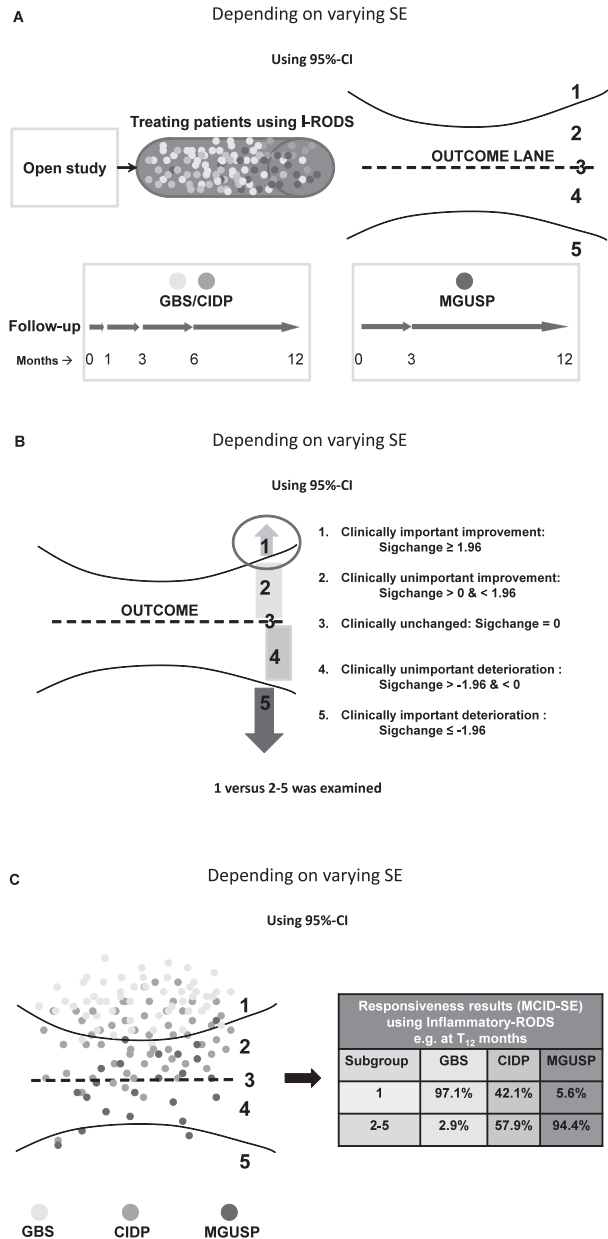
$$\text{Sigchange} = T_i - T_0 / \sqrt{(SE_{T_i}^2 + SE_{T_0}^2)}$$

- Cut-off used for comparison purposes: 1.96 x SE (95% certainty)

Subgroup 1. Clinically important improvement:	Sigchange ≥ 1.96
Subgroup 2. Clinically unimportant improvement:	Sigchange > 0 & < 1.96
Subgroup 3. Clinically unchanged:	Sigchange = 0
Subgroup 4. Clinically unimportant deterioration :	Sigchange > -1.96 & < 0
Subgroup 5. Clinically important deterioration :	Sigchange ≤ -1.96

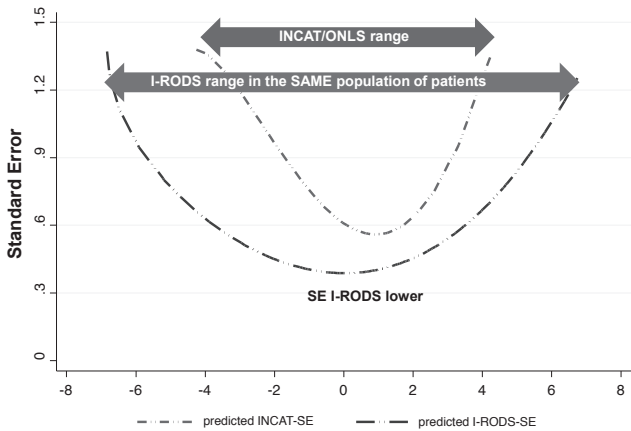
**Legend to Figure e-1:** A: Schematic representation of the U-shape SE-pattern across an outcome measure of interest; B: formula used to define important improvement being defined as ≥ 1.96 x SE for each individual patient, at a chosen of 95% confidence interval.<sup>20,21</sup>  
SE: standard error.

**Figure e-2.** Explaining the use of the individually obtained Standard Errors (SE) through Rasch determining “being a responder”



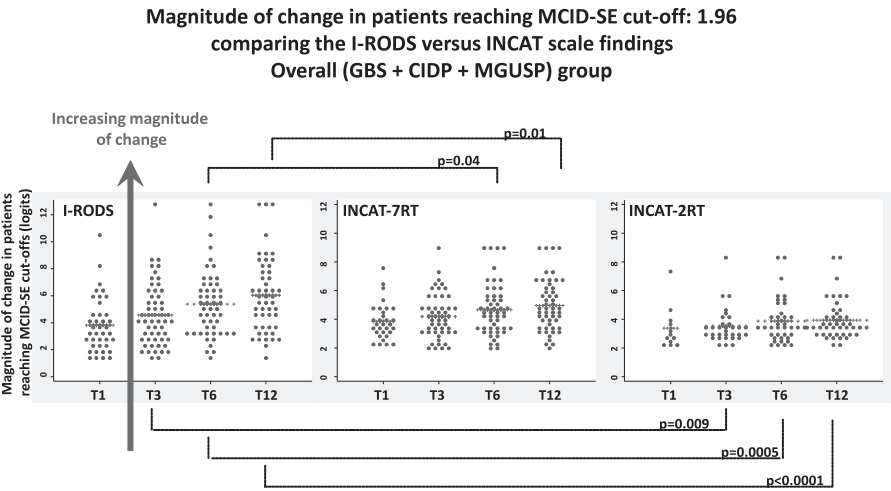
**Legend to Figure e-2:** From the formula described in Figure e-1B, 5 subgroups of patients were determined: a cut-off  $\geq 1.96 \times \text{SE}$  was defined as being a responder and applied for all comparison studies as part of the PeriNomS study.<sup>20,21</sup> The figure should be read from A→B→C. SE: standard error; CI: confidence interval; I-RODS: inflammatory Rasch-built overall disability scale.

**Figure e-3.** Location on the I-RODS versus INCAT-ONLS related to the corresponding individual standard errors



**Legend to Figure e-3:** The range of the I-RODS was larger in the same population at examination compared to the findings in the INCAT-ONLS scales. Data shown here are for the RT-7-item INCAT-ONLS scale. The corresponding SEs were practically the same for the RT-2-item (arm and leg grade) INCAT-ONLS as for the RT-7-item INCAT-ONLS. I-RODS: inflammatory Rasch-built Overall Disability Scale; INCAT: Inflammatory Neuropathy Cause And Treatment Scale; ONLS: overall neuropathy limitation scale; RT: Rasch transformed; SE: standard error.

**Figure e-4:** Magnitude of change in patients reaching MCID-SE cut-off: 1.96: Comparing the I-RODS versus INCAT findings, Overall (GBS + CIDP + MGUSP) group.



**Legend to Figure e-4:** T1: assessment at 1 month, T3: at 3 months, T6: at 6 months, T12: at 12 months. I-RODS: inflammatory Rasch-built overall disability scale; INCAT-7RT: INCAT 7-item Rasch-transformed scale, INCAT-2RT: INCAT 2-item Rasch-transformed scale. Student's T-test was used for comparison purposes. GBS: Guillain-Barré syndrome; CIDP: chronic inflammatory demyelinating polyradiculoneuropathy; MGUSP: monoclonal gammopathy of undetermined significance related polyneuropathy. +++: mean values.



# Defining a responder through Rasch analyses

*Els K. Vanhoutte<sup>1</sup>, MD; Catharina G. Faber<sup>1</sup>, MD, PhD; Sonja I. van Nes<sup>2</sup>, MD, PhD; Kenneth C. Gorson<sup>3</sup>, MD; P.A. van Doorn<sup>2</sup>, MD, PhD; David R. Cornblath<sup>4</sup>, MD, PhD; Ingemar S.J. Merkies<sup>1,5</sup>, MD, PhD;  
on behalf of the PeriNomS study group*

<sup>1</sup>Department of Neurology, University Medical Centre Maastricht, Maastricht, the Netherlands;

<sup>2</sup>Department of Neurology, Erasmus Medical Centre Rotterdam, Rotterdam, the Netherlands;

<sup>3</sup>Department of Neurology, St. Elizabeth's Medical Center, Tufts University School of Medicine, Boston, USA; <sup>4</sup>Department of Neurology, Johns Hopkins School of Medicine, Baltimore, USA;

<sup>5</sup>Department of Neurology, Spaarne Hospital, Hoofddorp, the Netherlands.

*Submitted*



## Abstract

This study aimed to define a “responder” through the concept of Rasch-determined minimum clinically important differences using the obtained individual SE's (MCID-SE) in patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and IgM-gammopathy related neuropathy (MGUSP). One hundred and thirty-seven newly diagnosed or clinically relapsing patients (GBS: 55, CIDP: 59, IgM-MGUSP: 23) were serially examined during 1-year follow-up. The following outcome measures were assessed: inflammatory neuropathy Rasch-built overall disability scale (I-RODS), Rasch-transformed MRC sum-score (RT-MRC), and Rasch-transformed modified INCAT sensory scale (RT-mISS). A MCID-SE cut-off  $\geq 1.96$  was defined as significant improvement, and the patient was then considered a responder.

In all three diseases, the SE values showed a characteristic “U”-shaped dynamic pattern across the scales' continuum. The number of patients showing a meaningful change were higher for the I-RODS > RT-MRC > RT-mISS and were higher in GBS compared to CIDP patients. There was virtually no improvement in patients with MGUSP.

The MCID-SE concept using Rasch-transformed data demonstrated a dynamic pattern of ‘being-a-responder’ in patients with immune-mediated neuropathies. The I-RODS scale showed greater responsiveness compared to other scales. Future treatment trials in these conditions should incorporate the MCID-SE to define a responder.

## Introduction

Clinical trials in patients with immune-mediated neuropathies like Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and monoclonal gammopathy of undetermined significance related neuropathy (MGUSP) have traditionally used mean differences or predefined, arbitrarily chosen cut-offs of ordinal constructed outcome measures in order to determine a significant difference between the placebo and the intervention group.<sup>1-8</sup> However, a statistically significant difference between treatment and placebo does not necessarily provide clinically meaningful information about the relevance of observed changes. Also, a statistical difference often depends on the nature of the selected outcome measure and on the sample size.<sup>9</sup> For example, in studies with very large sample sizes, small differences may be statistically significant but not necessarily clinically meaningful.<sup>10</sup>

Studies have shown that the clinical relevance of a treatment effect can be determined by using the concept of a “minimum clinically important difference” (MCID).<sup>11,12</sup> The MCID can be determined by either “anchor-based” or “distribution-based” methods.<sup>13,14</sup> In anchor-based techniques the patient indicates whether or not improvement has been achieved.<sup>13</sup> Distribution-based methods are based on a statistical approach, e.g., using a cut-off based on the standard deviation of the theoretical total spread of an outcome measure of interest or the standard error of the mean.<sup>13,15,16</sup> The efficacy of intravenous immunoglobulin (IVIg) compared to placebo was demonstrated in patients with CIDP using the MCID.<sup>17</sup> However, defining a responder using a “fixed cut-off MCID” implies a fixed standard error (SE) estimate across the range of the outcome measure used.<sup>11</sup> However, modern clinimetric methods such as Rasch analysis have shown variable SE values related to the varying ability (location) of a patient throughout a scale’s continuum, with lower SE scores in the mid-section of patient’s ability, and increasing SE values towards the extremes of a patient’s ability.<sup>18</sup> Thus far no treatment study has ever used the varying SE observed across the continuum of any outcome scale in patients with GBS, CIDP, and MGUSP. We determined the variable SE across several commonly used outcome measures applied in these conditions after converting the ordinal data to interval measures using the Rasch method,<sup>19,20</sup> and defined a treatment responder through the concept of MCID-SE in serially examined patients.

## Methods

### Patients, approvals, and patient consents

This study was conducted as part of the Peripheral Neuropathy Outcome Measures Standardization (PeriNomS) study (duration: 7 years; database closed 31<sup>st</sup> December 2012), an international collaborative effort of 26 neuromuscular centres from nine different countries (USA, Canada, Spain, Brazil, Italy, France, Belgium, the Netherlands,

and UK) with a special interest in immune-mediated neuropathies.<sup>18,21</sup> The primary aims of the PeriNomS study were to select the most responsive outcome measures in GBS, CIDP, and MGUSP, and to develop a core set of outcome measures assessing impairment, activity and participation, and quality of life for each illness to be used in future studies. One hundred and thirty-seven patients with newly diagnosed GBS or clinical relapsing CIDP, and IgM MGUSP were recruited by the participating centres and were serially examined during a 12-month period. Patients received treatment at the discretion of their physician. Eligibility included normal cognitive function, age  $\geq 18$  years, fulfilling international criteria for the diagnosis GBS, CIDP, or IgM MGUSP (most with anti-myelin-associated-glycoprotein antibodies (MAG)+),<sup>22-24</sup> having progressive impairment or disability (GBS) or a clinical relapse (CIDP or MGUSP) as judged by the patient and physician, and no treatment for their polyneuropathy for at least 2 months before inclusion. Patients were excluded if other causes of polyneuropathy were present (e.g. renal insufficiency, diabetes mellitus, vitamin deficiencies, etc.) or any other conditions (e.g. alcohol abuse:  $> 5$  IU/day) that might interfere with nervous system and physical functioning (e.g. arthritis). All patients gave written informed consent before participation in the study.

### Standard protocol approvals, registrations, and patient consents

The local Medical Ethics Committee in each participating center approved the protocol. Written informed consent was obtained from all participants.

### Outcome measures applied

The selection of outcome measures was based on recommendations given at the workshop on outcome measures in inflammatory neuropathies.<sup>21</sup>

The following impairment and disability measures were selected:

- The *MRC sum score* comprises the following six muscle pairs: upper arm abductors, elbow flexors, wrist extensors, hip flexors, knee extensors, and foot dorsal flexors.<sup>25</sup> These muscles were initially examined and scored using the Medical Research Council grades that range from 0 (paralysis) to 5 (normal strength).<sup>26</sup>
- The *Modified INCAT sensory scale (mISS)* assesses light touch, pinprick, vibration, and joint position sense in the arms and legs, and two-point discrimination (using a sliding aesthesiometer) at the index finger.<sup>27</sup> Except for two-point discrimination, all sensory modalities were scored separately using response options that ranged from 0 (no sensory deficit at the index finger or hallux) to 4 (highest sensory deficit at acromioclavicular joint or anterior superior iliac spine). Two-point discrimination was scored as normal (grade 0) or disturbed (grade 1) based on age-dependent normative values.<sup>28</sup>
- The inflammatory *Rasch-built overall disability scale for GBS, CIDP and MGUSP (I-RODS)* consists of 24 daily and social activity items that are scored as 0 (impossible to perform), 1 (performed with difficulty), or 2 (easily performed, without difficulty).<sup>29</sup>

This outcome measure at the interval level was specifically designed to capture activity limitations and participation restrictions in patients with GBS, CIDP, and MGUSP and proved to be superior to the widely used ordinal INCAT disability scale and related measures (ODSS and ONLS).<sup>18,30-32</sup>

## Study procedures

As part of the PeriNomS study, all centers participated in standardized training of the assessment procedures for all the outcome measures. In addition, participants received a research manual that integrated a thorough description as well as pictures illustrating how to assess the MRC grades and the various sensory modalities for the mISS. Patients received standardized written instructions for the completion of the I-RODS.

## Rasch analyses and statistical aspects

- **Creating interval measures through Rasch**

The Rasch methodology has been described in detail elsewhere.<sup>17,19,29,33</sup> The data for all selected outcome measures were subjected to Rasch analysis using the software Rasch unidimensional measurement model (RUMM2030).<sup>34</sup> The RUMM2030 software enables the transformation of ordinal obtained scores into interval measures, and provides an estimate of each patient's location with a corresponding SE on the same logit (log-odds unit) scale.<sup>35,36</sup>

- **Determining MCID-SE based on the clinical picture**

Various methods have been proposed to capture clinical relevant changes in patients.<sup>14,17,37</sup> However, there is no consensus on how to quantify the importance and magnitude of such changes.<sup>17</sup> Also, no study in inflammatory neuropathies has considered the possibility that the magnitude of clinically relevant change may vary, dependent on the clinical picture of the diagnosis and variable SE corresponding to the ability of the patient.<sup>18,35</sup>

In GBS and CIDP, the temporal course is acute or subacute and characterized by a motor predominant neuropathy. Sensory features are common but may contribute less to functional disability.<sup>38</sup> Therefore, weakness assessed by the MRC sum-score may have greater importance related to disability than sensory changes as measured by the mISS. In contrast, MGUSP tends to be more indolent with less impairment and disability, at least early in the course.<sup>24</sup> Sensory abnormalities are often modest and more common than weakness, and therefore outcome measures may be less responsive over short time periods in this condition.

- **Responsiveness at the individual person level**

Minimal Clinically Important Difference-Standard Error (MCID-SE) score based on previously described significant change (SigChange).<sup>35</sup>

In brief, MCID-SE was calculated by computing for each participating patient at each assessment (GBS/CIDP: T0, T1, T3, T6, and T12; MGUSP: T0, T3, and T12 months) and for each scale applied:

- their own change (person location at  $i$  month minus person location at entry; where  $i = 1, 3, 6$  or 12 months of follow-up in patients with GBS and CIDP or 3 or 12 months in patients with MGUSP),
- the corresponding SE of difference related to their individual change ( $SE_{diff} = \text{square-root}(SE_{entry}^2 + SE_{i\text{ month}}^2)$ ), and
- The final MCID-SE was calculated by dividing the individual change in scores by corresponding  $SE_{diff}$ :  $MCID-SE = (\text{person location at } i \text{ month minus person location at entry})/SE_{diff}$ .

The degree of uncertainty associated with the ability estimate for each individual patient was expressed by the SE and was used to calculate the confidence intervals around the unknown true ability score. Accordingly, 1 SE provides 68% certainty that the ability estimation was within this range; 1.96 SE provided a 95% confidence interval.<sup>18,35,36</sup> Meaningful change was defined as an estimated ability score  $\geq 1.96$  SE.<sup>35</sup> This MCID-SE cut-off was used for the purposes of the current study for all outcome measures to determine meaningful changes. In addition, in MGUSP patients an MCID-SE cut-off of 1 SE also was calculated because of the likelihood of less responsiveness of the scales due to the indolent nature of this illness.

Five subgroups were determined, based on the MCID-SE scores:

- subgroup 1 (clinically important improvement):  $MCID-SE \geq 1.96$ ;
- subgroup 2 (clinically unimportant improvement):  $0 < MCID-SE < 1.96$ ;
- subgroup 3 (no change):  $MCID-SE = 0$ ;
- subgroup 4 (clinically unimportant deterioration):  $-1.96 < MCID-SE < 0$ ;
- subgroup 5 (clinically important deterioration):  $MCID-SE \leq -1.96$ .

In patients with MGUSP, similar subgroups also were calculated using  $MCID-SE \geq 1$ . The MCID-SE cut-offs were reversed for the mISS, since a reduction in this scale's score reflects improvement.

#### • Software

Rasch analyses were performed with the partial credit model as default (RUMM2030).<sup>34</sup> Graphs were constructed using Stata Statistical Software for Windows XP (version 12.0, StataCorp, Texas, USA).

## Results

### Patients

The demographics of the 137 patients are presented in Table 1. Anti-MAG antibodies were present in 19 of the 23 patients with IgM MGUSP. There were 600 completed visits and 94% of the patients completed a 1-year follow-up.

**Table 1.** General characteristics of n=137 longitudinal recruited patients

Diagnosis à	GBS	CIDP	MGUSP
# patients recruited →	55	59	23
<b>Age (years)</b>			
mean (SD), range	53.4 (17.7), 19-90	52.7 (14.7), 18-79	62.8 (10.6), 45-85
<b>Gender, # (%)</b>			
Female	18 (32.7)	17 (28.8)	8 (34.8)
Male	37 (67.3)	42 (71.2)	15 (65.2)
<b>Country of assessment</b>			
The Netherlands	25	20	8
USA	6	11	4
Italy	9	9	3
Spain	-	2	-
France	1	4	1
Belgium	2	2	5
UK	2	2	1
Canada	8	7	1
Brazil	2	2	-
<b>Current medication (%)</b>			
IVIg and/or Steroids	81.8%	67.8%	34.8%
Plasma Exchange	16.4%	5.1%	17.4%
Immunosuppressive	-	3.4%	13%
Other (including no therapy)	-	10.2%	34.8%

**Variable SE with changing patients’ location**

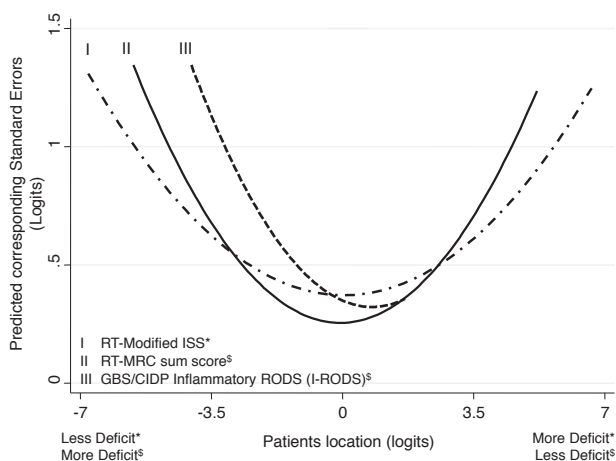
Rasch transformation of the MRC summed scores and mISS (RT-MRC; RT-mISS) created interval measures with corresponding locations and SE for each patient. In order to preserve the original structure of both scales, factors such as item bias, local dependency, and good fit statistics were not corrected.

Figure 1 illustrates the variable SE with changing patient locations on the Rasch-built outcome measures for all the patients. A “U-shaped” pattern was observed, with lower SE scores in the mid-section of patients’ ability and increasing SE values at the extremes of patients’ ability in all metrics applied. The highest variation in SE was seen in the RT-MRC (5.2-fold); the SE varied 4.4-fold for the RT-mISS and 3.5-fold for the I-RODS. The range of the I-RODS was much broader than the other two measures.

**Determining clinical relevant changes**

Tables 2 and 3 and Figure 2 provide the MCID-SE category results of all individual patients depending on the magnitude of the clinical changes and the direction (deterioration, stable, improvement) of these changes. Overall, patients with GBS demonstrated a much faster and consistently higher percentage of clinically important improvement when compared to patients with other inflammatory neuropathies. The highest responsiveness was seen for the I-RODS scale (97.1% responders at 12 months), followed by the Rasch-modified MRC sumscore (70.6% responders at 12 months). There was virtually no clinically important improvement in patients with MGUSP using a cut-off of  $\geq 1.96 \times \text{SE}$ .

**Figure 1.** The ‘U-shape’ of the variable standard errors (SE) with changing patient locations on the various interval outcome measures applied in patients with immune-mediated neuropathies.



**Legend to Figure 1:** GBS = Guillain-Barré syndrome, CIDP = chronic inflammatory demyelinating polyradiculoneuropathy, MGUSP = monoclonal gammopathy of undetermined significance related neuropathy, SE = standard error. MRC = Medical Research Council, mISS = modified INCAT sensory sumscore, I-RODS = Rasch-built overall disability scale for GBS, CIDP and MGUSP. RT: Rasch-Transformed

**Table 2.** Categorizing the minimum clinically important difference using the variable SE values (MCID-SE  $\geq 1.96$ ) for each longitudinally examined patient with GBS and CIDP according to the size and direction of changes seen for all outcome measures applied

MCID-SE cut-off value: 1.96		GBS (n=55)											
classification (%) based on MCID-SE values		I-RODS				RT-MRC sumscore				RT-modified ISS			
Follow-up period (months) $\Delta$		1	3	6	12	1	3	6	12	1	3	6	12
1: Clinically important improvement		48.8	78.1	89.5	97.1	36.4	58.5	71.1	70.6	22.7	34.2	47.4	35.3
2: Clinically unimportant improvement		37.2	17.1	7.9	0	38.6	29.3	18.4	20.6	47.7	43.9	34.2	44.1
3: Stable (unchanged)		2.3	2.4	0	0	9.1	7.3	7.9	8.8	13.6	12.2	13.2	14.7
4: Clinically unimportant deterioration		9.3	2.4	2.6	0	11.4	4.9	2.6	0	13.6	9.8	5.3	5.9
5: Clinically important deterioration		2.3	0	0	2.9	4.5	0	0	0	2.3	0	0	0
MCID-SE cut-off value: 1.96		CIDP (n=59)											
classification (%) based on MCID-SE values		GBS/CIDP RODS				RT-MRC sumscore				RT-modified ISS			
Follow-up period (months) $\Delta$		1	3	6	12	1	3	6	12	1	3	6	12
1: Clinically important improvement		29.8	42.2	46.2	42.1	19.2	34.1	33.3	29	10.9	20.9	15.8	21.6
2: Clinically unimportant improvement		40.4	35.6	35.9	29	42.5	27.3	38.5	44.7	50	39.5	63.2	54.1
3: Stable (unchanged)		12.8	4.4	7.7	5.3	14.9	25	15.4	10.5	19.6	18.6	5.3	8.1
4: Clinically unimportant deterioration		10.6	13.3	7.7	18.4	19.1	11.4	12.8	13.2	17.4	20.9	15.8	16.2
5: Clinically important deterioration		6.4	4.4	2.6	5.3	4.3	2.2	0	2.6	2.1	0	0	0

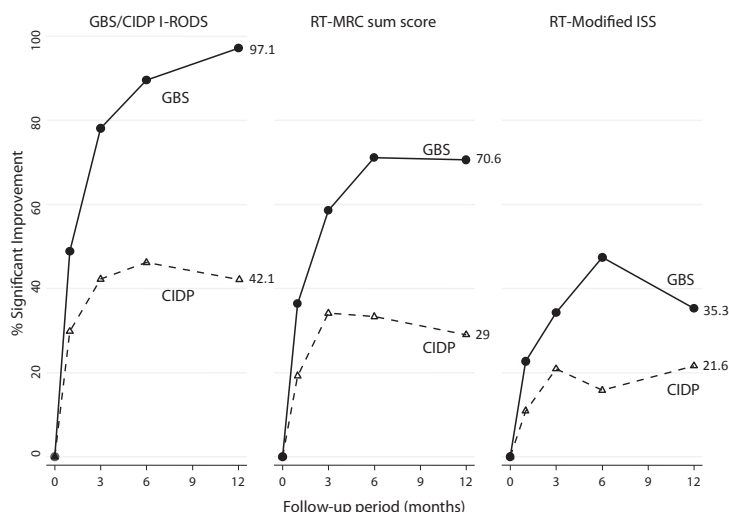
**Legend to Table 2:** Minimum clinically important difference using varying standard errors (MCID-SE); GBS = Guillain-Barré syndrome, CIDP = chronic inflammatory demyelinating polyradiculoneuropathy, MGUSP = monoclonal gammopathy of undetermined significance related neuropathy. SE = standard error. MRC = medical Research Council, mISS = modified INCAT sensory sumscore, I-RODS = Inflammatory Rasch-built overall disability scale for GBS, CIDP and MGUSP. RT: Rasch-Transformed. The shaded cells reflect clinically important improvement.

**Table 3.** Categorizing the minimum clinically important difference using the variable SE values (MCID-SE) of each longitudinally examined patient with MGUSP according to the size and direction of changes seen for all outcome measures applied, A: MCID-SE cut-off value at 1x SE and B: cut-off value at 1.96x SE

Subgroup classification (%) of all patients examined serially based on their MCID-SE values	MGUSP (n=23)					
	I- RODS	RT-MRC sumscore	RT-modified ISS			
A: MCID-SE cut-off value: 1						
Follow-up period (months) →	3	12	3	12	3	12
1: Clinically important improvement	35	27.8	20	17.7	25	17.7
2: Clinically unimportant improvement	15	27.8	15	17.7	35	41.2
3: Stable (unchanged)	15	11.1	50	47.1	25	11.8
4: Clinically unimportant deterioration	25	11.1	10	0	5	17.7
5: Clinically important deterioration	10	22.2	5	17.7	10	11.7
B: MCID-SE cut-off value: 1.96						
Follow-up period (months) →	3	12	3	12	3	12
1: Clinically important improvement	20	5.6	5	5.9	10	5.9
2: Clinically unimportant improvement	30	50	30	29.4	50	52.9
3: Stable (unchanged)	15	11.1	50	47.1	25	11.8
4: Clinically unimportant deterioration	35	27.8	10	5.9	15	29.4
5: Clinically important deterioration	0	5.6	5	11.8	0	0

**Legend to Table 3:** Minimum clinically important difference using varying standard errors (MCID-SE); MGUSP = monoclonal gammopathy of undetermined significance related neuropathy. SE = standard error. MRC = medical Research Council, mISS = modified INCAT sensory sumscore, I-RODS = Inflammatory Rasch-built overall disability scale for GBS, CIDP and MGUSP. RT: Rasch-Transformed. The shaded cells reflect clinically important improvement.

**Figure 2.** Percentage of GBS and CIDP patients reaching clinically important changes applying the concept of minimum clinically meaningful changes using different cut-offs (MCID-SE  $\geq 1.96$ ) in various interval outcome measures



**Legend to Figure 2:** Patients with GBS consistently demonstrated higher scores when compared to patients with CIDP. The I-RODS demonstrated the highest percentages of responders in both GBS and CIDP. MCID-SE = minimum clinically important difference using varying individual standard errors (MCID-SE). GBS = Guillain-Barré syndrome, CIDP = chronic inflammatory demyelinating polyradiculoneuropathy, MRC = Medical Research Council, mISS = modified INCAT sensory sumscore, I-RODS = Inflammatory Rasch-built overall disability scale for GBS, CIDP and MGUSP. RT: Rasch-Transformed



## Discussion

We have demonstrated that SE values vary depending upon the patient's level of impairment and the difficulty of the activity in question after subjecting frequently used impairment and disability outcome measures in patients with inflammatory neuropathies to Rasch analysis. These variable SE values enable the definition of a "responder" at the individual person level (patient-targeted approach), and may therefore have major implications for future trials in these conditions by defining a clinically meaningful change for any of the interval outcome measures we studied. The concept of MCID-SE captures clinically meaningful improvement as well as the proportion of patients showing meaningful deterioration. The MCID-SE is a **one-stop design** that captures changes in all directions and may therefore be of particular importance for maintenance trials as well studying new therapeutic agents in previously treated patients with these conditions.

Although in many clinical trials ordinal outcome measures (with Likert-type response options) have been analyzed as if the responses were a linear measure, the true distance between the different response categories remains unknown, thereby making calculations with sum scores less meaningful.<sup>39,40</sup> In addition, these ordinal outcome measures are used to calculate a group level SE that is assumed to remain equal for all patients regardless their changing ability level. Modern clinimetric methods like Rasch analysis transforms ordinal data into interval measurement and calculates the precise SE for each person location. Our findings using the Rasch method have demonstrated that the SE is related to the ability level of the patients, thereby varying across the range of the scale (Figure 1).<sup>35,41,42</sup>

Furthermore, the natural course of the disease needs to be taken into account when establishing a cut-off value for clinically important change. For example, the high percentage of GBS and CIDP patients who reached the 1.96 MCID-SE cut-off for the RODS indicated that this outcome measure captures the dynamic changes that occur during the course of treatment in these illnesses. In addition to good validity, reliability, and fulfillment of the Rasch model expectations, we showed that the I-RODS has the greatest responsiveness in GBS and CIDP, when using the MCID-SE method for defining a responder.<sup>18,29</sup> Accordingly, we propose that the I-RODS should be the preferred outcome scale to assess activity limitations and participation restrictions in patients with GBS and CIDP.<sup>18</sup> There were only a few responders in patients with MGUSP at both cut-offs (MCID-SE 1, and  $\geq 1.96$ ) in all the scales that were used, perhaps due to the indolent nature of the disease. These findings in patients with MGUSP suggest that assessment in this condition may require a period longer than 12 months or the development of even more sensitive outcome measures specifically for this illness.

There are some methodological issues that should be addressed. Although the MRC sumscore and mISS data were analyzed using Rasch analysis, the original data were incompletely transformed to meet Rasch model requirements, as factors such as item

bias, local dependency, and fit statistics were not corrected.<sup>33</sup> Our aim was to keep the structure of these ordinal scales as complete as possible, but for comparison purposes we had to subject the measures to RUMM2030, in order to create a higher level of assessment precision at the interval level as is the case for the I-RODS. However, we have previously shown that physicians are not able to discriminate between the six MRC response categories for most muscle groups, with disordered thresholds in most examined muscles.<sup>43,44</sup> Transforming the MRC six grades to four categories would lead to higher individual SEs, as we recently have demonstrated.<sup>43</sup> We would expect the responsiveness values for the individual patient to decrease to a more “accurate” level using the Rasch-transformed MRC sum score four grade scale using the MCID-SE formula ( $\text{MCID-SE} = \text{person location at } i \text{ month} - \text{person location at entry} / \text{SE}_{\text{diff}}$  where  $\text{SE}_{\text{diff}} = \text{square-root} (\text{SE}_{\text{entry}}^2 + \text{SE}_{i \text{ month}}^2)$ ). In essence, the proportion of patients with a clinically meaningful response would be lower, since the denominator in the equation increases with the increasing SEs (data not shown). A similar finding would be expected if the mISS was transformed to meet Rasch model expectations. Despite this, our findings in the current study show a lower proportion of patients with a clinically important improvement using the MRC and mISS scales when compared to the I-RODS. Lastly, it remains controversial what estimated score change should be considered a clinically meaningful change. In the absence of a consensus on this subject,<sup>14</sup> we suggest using the MCID-SE cut-off of 1.96, since it reduces the imprecise measurements seen in our estimated scores, taking the varying SEs into account. In summary, the current study shows that the standard error (SE) is not fixed but varies across the range of an outcome measure enabling the calculation of individual change using the concept of MCID-SE. Similar findings have been reported by others.<sup>35,41</sup> The I-RODS demonstrated lower variation in the SEs and had a broader range when compared to the MRC and mISS. These characteristics indicate that this scale is more responsive in patients with immune-mediated neuropathies when compared to the others studied.

## References

- 1 van Nes, S. I., Faber, C. G., and Merkies, I. S. Outcome measures in immune-mediated neuropathies: the need to standardize their use and to understand the clinimetric essentials. *J Peripher Nerv Syst* **13**, 136-147 (2008).
- 2 Allen, D., Lunn, M. P., Niermeijer, J. *et al.* Treatment for IgG and IgA paraproteinaemic neuropathy. *Cochrane Database Syst Rev*, CD005376 (2007).
- 3 Hughes, R. A., Swan, A. V., and van Doorn, P. A. Intravenous immunoglobulin for Guillain-Barre syndrome. *Cochrane Database Syst Rev*, CD002063 (2010).
- 4 Lunn, M. P. and Nobile-Orazio, E. Immunotherapy for IgM anti-myelin-associated glycoprotein paraprotein-associated peripheral neuropathies. *Cochrane Database Syst Rev*, CD002827 (2006).
- 5 Eftimov, F., Winer, J. B., Vermeulen, M. *et al.* Intravenous immunoglobulin for chronic inflammatory demyelinating polyradiculoneuropathy. *Cochrane Database Syst Rev*, CD001797 (2009).
- 6 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet Neurol* **7**, 136-144 (2008).
- 7 van Schaik, I. N., Eftimov, F., van Doorn, P. A. *et al.* Pulsed high-dose dexamethasone versus standard prednisolone treatment for chronic inflammatory demyelinating polyradiculoneuropathy (PREDICT study): a double-blind, randomised, controlled trial. *Lancet Neurol* **9**, 245-253 (2010).
- 8 Nobile-Orazio, E., Cocito, D., Jann, S. *et al.* Intravenous immunoglobulin versus intravenous methylprednisolone for chronic inflammatory demyelinating polyradiculoneuropathy: a randomised controlled trial. *Lancet Neurol* **11**, 493-502 (2012).
- 9 Sterne, J. A. and Davey Smith, G. Sifting the evidence-what's wrong with significance tests? *BMJ* **322**, 226-231 (2001).
- 10 Neely, J. G., Karni, R. J., Engel, S. H. *et al.* Practical guides to understanding sample size and minimal clinically important difference (MCID). *Otolaryngol Head Neck Surg* **136**, 14-18 (2007).
- 11 Nunnally, J. C. *Psychometric theory*. (McGraw Hill, 1978).
- 12 Jaeschke, R., Singer, J., and Guyatt, G. H. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* **10**, 407-415 (1989).
- 13 Copay, A. G., Subach, B. R., Glassman, S. D. *et al.* Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* **7**, 541-546 (2007).
- 14 Beaton, D. E., Boers, M., and Wells, G. A. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol* **14**, 109-114 (2002).
- 15 Sloan, J. A., Cella, D., Frost, M. *et al.* Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clin Proc* **77**, 367-370 (2002).
- 16 Norman, G. R., Sloan, J. A., and Wywich, K. W. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* **41**, 582-592 (2003).
- 17 Merkies, I. S., van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *J Neurol Neurosurg Psychiatry* **81**, 1194-1199 (2010).
- 18 Vanhoutte, E. K., Faber, C. G., Merkies, I. S. *et al.* 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscul Disord* **23**, 924-933 (2013).
- 19 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1960).
- 20 Lin, K. C., Fu, T., Wu, C. Y. *et al.* Assessing the stroke-specific quality of life for outcome measurement in stroke rehabilitation: minimal detectable change and clinically important difference. *Health Qual Life Outcomes* **9**, 5 (2011).
- 21 Merkies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 22 Van der Meche, F. G., Van Doorn, P. A., Meulstee, J. *et al.* Diagnostic and classification criteria for the Guillain-Barre syndrome. *European neurology* **45**, 133-139 (2001).

- 23 Joint Task Force of the EFNS and the PNS. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of chronic inflammatory demyelinating polyradiculoneuropathy: report of a joint task force of the European Federation of Neurological Societies and the Peripheral Nerve Society--First Revision. *J Peripher Nerv Syst* **15**, 1-9 (2010).
- 24 Joint Task Force of the EFNS and the PNS. European Federation of Neurological Societies/Peripheral Nerve Society Guideline on management of paraproteinemic demyelinating neuropathies. Report of a Joint Task Force of the European Federation of Neurological Societies and the Peripheral Nerve Society--first revision. *J Peripher Nerv Syst* **15**, 185-195 (2010).
- 25 Kleyweg, R. P., van der Meche, F. G., and Schmitz, P. I. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle Nerve* **14**, 1103-1109 (1991).
- 26 Medical Research Council. Medical Research Council. Aids to the investigation of the peripheral nervous system. 1-2 (Her Majesty's Stationary Office, London, 1943).
- 27 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 28 van Nes, S. I., Faber, C. G., Hamers, R. M. *et al.* Revising two-point discrimination assessment in normal aging and in patients with polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **79**, 832-834 (2008).
- 29 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 30 Hughes, R., Bensa, S., Willison, H. *et al.* Randomized controlled trial of intravenous immunoglobulin versus oral prednisolone in chronic inflammatory demyelinating polyradiculoneuropathy. *Annals of neurology* **50**, 195-201 (2001).
- 31 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Clinimetric evaluation of a new overall disability scale in immune mediated polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **72**, 596-601 (2002).
- 32 Graham, R. C. and Hughes, R. A. A modified peripheral neuropathy scale: the Overall Neuropathy Limitations Scale. *Journal of neurology, neurosurgery, and psychiatry* **77**, 973-976 (2006).
- 33 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 34 Andrich, D., Sheridan, B., and Luo, G. *Rasch models for measurement: RUMM2030*. (2010).
- 35 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* **13**, iii, ix-x, 1-177 (2009).
- 36 Wright, B. D. and Stone, M. H. *Best test design*. (Mesa Press, 1979).
- 37 Liang, M. H. Evaluating measurement responsiveness. *J Rheumatol* **22**, 1191-1192 (1995).
- 38 Whitesell, J. Inflammatory neuropathies. *Semin Neurol* **30**, 356-364 (2010).
- 39 Merbitz, C., Morris, J., and Grip, J. C. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil* **70**, 308-312 (1989).
- 40 Hobart, J. C., Cano, S. J., Zajicek, J. P. *et al.* Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol* **6**, 1094-1105 (2007).
- 41 Heesch, K. C., Masse, L. C., and Dunn, A. L. Using Rasch modeling to re-evaluate three scales related to physical activity: enjoyment, perceived benefits and perceived barriers. *Health Educ Res* **21 Suppl 1**, i58-72 (2006).
- 42 van der Beek, N. A., Hagemans, M. L., van der Ploeg, A. T. *et al.* The Rasch-built Pompe-specific activity (R-PAct) scale. *Neuromuscul Disord* **23**, 256-264 (2013).
- 43 Vanhoutte, E. K., Faber, C. G., van Nes, S. I. *et al.* Modifying the Medical Research Council grading system through Rasch analyses. *Brain* **135**, 1639-1649 (2012).
- 44 Shaw, F., Wright, B., and Linacre, J. M. Disordered steps? *Rasch Measurement Transactions* **6**, 225 (1992).



## **Part IV**

# **Summary and future perspectives**



## Chapter 12

### Workshop report

196th ENMC International Workshop:

# **'Outcome Measures in inflammatory peripheral Neuropathies'**

**8 – 10 February 2013, Naarden, the Netherlands**

*Els K. Vanhoutte<sup>a</sup>; Catharina G. Faber<sup>a</sup>; Ingemar S.J. Merkies<sup>ab</sup>; on behalf of the PeriNomS study group*

<sup>a</sup>Maastricht University Medical Centre, Maastricht, the Netherlands;

<sup>b</sup>Spaarne Hospital, Hoofddorp, the Netherlands

*Neuromuscul Disord. 2013 Nov;23(11):924-33*



## Introduction

For the current 196<sup>th</sup> ENMC international workshop, twenty researchers in the field of neuromuscular disorders from eight different countries (USA, Canada, Spain, Italy, France, Belgium, the Netherlands, UK), and a patient representative of the GBS/CIDP Foundation International (Glennys Sanders) participated in the workshop in Naarden, the Netherlands from February 8<sup>th</sup>-10<sup>th</sup> 2013 to discuss the results of the Peripheral Neuropathy Outcome Measurement Standardisation (PeriNoms) study and to strive for consensus on a specific core set of outcome measures to be used in future clinical trials and follow-up studies in patients with Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), monoclonal gammopathy of undetermined significance related polyneuropathy (MGUSP), and multifocal motor neuropathy (MMN).

In 2004, the first workshop on outcome measures in peripheral neuropathies was held on the auspice of the European Neuromuscular Centre (ENMC). At this 131<sup>st</sup> workshop, inadequate agreement was reached regarding which outcome measures should be used at the impairment and activity and participation levels in future trials involving patients with inflammatory neuropathies, like GBS, CIDP, MGUSP (IgM paraproteinemia with positive anti-myelin associated glycoprotein antibodies), and MMN.<sup>1</sup> Recommendations were provided to perform a comparison study between outcome measures that formed the basis of the current workshop. Based on these recommendations, the PeriNomS study was conducted between 2007 and 2012.

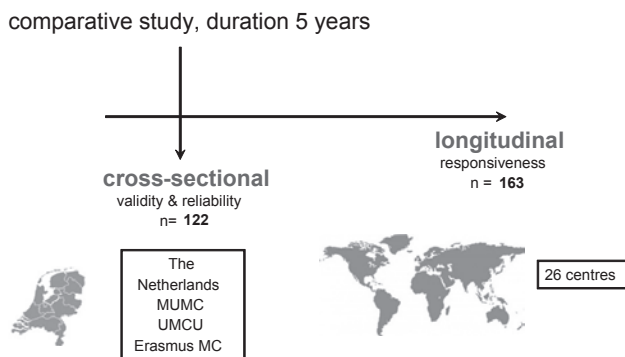
## Background

### PeriNomS study: background and recruitment findings

The PeriNomS study (2 years of preparation, 5 years of data collection; kick-off 2007, PNS, Utah, USA; dbase closed 31<sup>st</sup> December 2012) is an international collaborative effort of 26 neuromuscular centres with special interest in inflammatory neuropathies. Through comparative responsiveness studies, the most responsive measures were selected at the various levels of interest of assessing outcome. The study consisted of two parts: the cross-sectional part and the longitudinal part (Figure 1). The cross-sectional part mainly focused on examining validity and reliability in 122 recruited patients (GBS: 30, CIDP: 30, MGUSP: 20, MMN: 22, inflammatory small fiber neuropathy (SFN): 20) with a clinical stable condition with residual symptoms and signs. These patients were examined twice (interval 2-4 weeks) by two well trained and independent observers to obtain inter-/intra-observer reliability scores. Patient-reported outcomes were assessed twice by patients in the same way, without having access to the previous scores. The longitudinal studies were performed worldwide (at the collaborating centres) and the data were centrally collected (Erasmus MC and Maastricht UMC,

the Netherlands), and were used to determine comparative responsiveness scores between ~equally valid and reliable pre-selected outcome measures in inflammatory neuropathies. A total of 163 newly diagnosed patients (GBS: 55, CIDP: 59, MGUSP: 23 of which 19 were IgM anti-myelin associated glycoprotein positive, MMN: 26) were enrolled and examined during a one-year period (GBS and CIDP: examined at T0, T1, T3, T6, and T12 months; MGUSP and MMN: examined at T0, T3, and T12 months). Basic characteristics of these patients were presented including treatment regimens if reported. The workshop mainly focused on the cross-sectional validity and reliability and longitudinal responsiveness results by comparing the (modified) Medical Research Council (MRC) sum score versus Neuropathy Impairment Scale (NIS) motor subset, the modified Inflammatory Neuropathy Cause And Treatment (INCAT) sensory sum score versus NIS sensory subset, the Jamar dynamometer versus the Martin Vigorimeter, the INCAT disability 10-point scale, the overall disability sum score (ODSS), the overall neuropathy limitations scale (ONLS) versus the newly devised inflammatory Rasch-built overall disability scale (R-ODS).<sup>2,3</sup>

**Figure 1.** The design of the cross-sectional and the longitudinal part of the PeriNomS study.



### Historical background of clinical assessments in inflammatory neuropathies

Richard Hughes, UK reviewed the historical background of the development of outcome measures focussing on strength, sensation and disability, relying in part on the historical essay provided by Dyck and colleagues.<sup>4</sup> In 1917 Robert Lovett, a Boston orthopaedist introduced a system for scoring weakness by manual muscle testing, ranging from 1 being normal to 6 paralysis.<sup>4</sup> In 1942 the order of scoring was reversed by the Medical Research Council to provide a scale for scoring strength ranging from 0 for paralysis to 5 for normal.<sup>4</sup> The MRC grading system was developed to assess peripheral nerve injuries but it has become used throughout all kinds of neurological illnesses. It became apparent that the MRC grades have unequal widths with grade 4 covering a much wider range than the other grades. The MRC sum score was developed derived adding the scores of 6 pairs of muscles to try to represent the overall strength of a

patient. This composite measure has served through the years as an outcome measure in numerous trials involving patients with inflammatory neuropathies.<sup>5</sup> In parallel, the Mayo Clinic developed what came to be called the Neuropathy Impairment Score with more grades of severity for milder weakness than the MRC grading system: muscle weakness is scored as follows: 0 = normal; 1 = 25%; 2 = 50%; 3 = 75% weak; 3.25 = movement against gravity just possible; 3.5 = movement with gravity eliminated just possible; 3.75 = flicker; and 4 = paralyzed.<sup>4</sup> Alternative methods of scoring sensory impairment have been developed, the INCAT sensory sum score (ISS) in Europe and the Neuropathy impairment sensory score (NIS<sub>sens</sub>) at the Mayo clinic.<sup>4,6</sup> Disability measures have been developed for inflammatory neuropathies such as 7-point disability scale for GBS and the 10-point Inflammatory Neuropathy Cause And Treatment (INCAT) disability for chronic inflammatory neuropathies. The latter has been further modified into the Overall Disability sum score (ODSS) and finally the Overall Neuropathy Limitations Score (ONLS).<sup>3</sup>

### **Trial design and outcome models in GBS and other inflammatory neuropathies**

Pieter van Doorn, the Netherlands, discussed the issues in trial design and outcome models. For rare disorders like inflammatory neuropathies, it may be important to use advanced prognostic modelling, which enables researchers to reduce the required sample size, hereby considerably shortening the length of trials in these conditions. In order to obtain comparability between trials, the use of previously used outcome measures was proposed. However, this statement was debated, since persistent use of improper outcome measures may continue to jeopardize the results in the same way as seen in the past. Proper attention should be given to the choice of the best outcome measure, particularly choosing outcome measures that have been tested in terms of their clinimetric properties in the disease of interest and fulfilling modern requirements.<sup>7</sup> Also, time of assessment during the course of the illness of interest is important, since the course of inflammatory neuropathies may vary and with this possible improvement in patients.<sup>8</sup> Possible reasons for trials in the inflammatory field being negative were addressed highlighting the possibility of insensitive outcome measures contributing to this. Finally, the need for international collaboration in these rare disorders was emphasized.

### **Requirements for Outcome Measures: Traditional aspects**

Ingemar Merkies, the Netherlands, presented on behalf of David R. Cornblath (USA) the traditional basic requirements for outcome measures. Several types of data were discussed. Outcome measures at the nominal and ordinal level are descriptive and the collected data are qualitative without having a numerical value. Despite this general knowledge, we have been allocating numbers to descriptive response categories of each item and have been doing parametric analyses on calculated sum scores, assuming linearity.<sup>9</sup> Measurements at the interval or ratio level increase the level of

precision in the assessment of interest since the recruited data have a numerical value. The distance between the categories is known, and therefore meaningful calculations can be performed.

The International Classification of Functioning, Disability and Health (ICF) is an international framework classifying the consequences of a health condition.<sup>10</sup> The ICF was discussed addressing its domains: body functions and structures (limitation in one of these is addressed as impairment, e.g., paresis) and activities and participation (capturing daily activities and social functioning, e.g., dressing up, fulfill a job). Limitations at any of these levels may contribute to the concept of quality of life that embraces patients' reaction to the discrepancy between actual and expected achievements arising as a consequence of their illness.<sup>7</sup>

Outcome measures should be simple, valid, reliable and responsive.<sup>7</sup> The advantages and disadvantages of traditional 'classic test theory' ordinal based composite outcome measures were summarized.<sup>9</sup> In brief, outcome measures in inflammatory neuropathies are generally ordinal multi-items composite measures that recruit items arbitrarily. Sum scores are generally used, assuming that all items have equal weight, hence assuming a fixed unit and thus treating the data as if these were at the linear level. Additionally the differences obtained in a sum score throughout the range of the constructed sum score have been equally treated. All these assumptions are highly unlikely and hamper the interpretation of the results in/between clinical trials.<sup>9</sup>

## **Modern requirements for outcome measures**

### **Requirements for Outcome Measures: Modern aspects through Rasch analyses**

Catharina Faber, the Netherlands, explained the 'modern' clinimetric approach through the concept of the Rasch model.<sup>11,12</sup> This model is based on a logical assumption: a patient with a high ability (a term used for any particular trait of interest; e.g., mobility, or performing daily and social activities etcetera; in essence, a less ill patient) will have a higher probability of affirming a particular task compared to a patient with a lower ability (more ill patient).<sup>11</sup> The Rasch method enables researchers to transform ordinal obtained data into an interval metric, hereby increasing the level of precision. Both the ability of a patient and the difficulty of a task, having a logarithmic relation, are placed on the same ruler (with logits as the fixed unit).<sup>11</sup> Items and patients are ordered according to the Guttman scaling procedure. An example of ordering items in terms of their difficulty and patients in terms of their ability on the same metric was provided in lay terms. The RUMM (Rasch Unidimension Measurement Model) was presented explaining through various real data examples some of its model requirements (like fit statistics, disordered thresholds, local dependency, item bias, and proper targeting). Emphasis was put on the need to review data that are not fulfilling model expectations and if possible to modify ("clean-up") these to improve model fitting.

### **ACTIVLIM, a measure of activity limitations in children and adults with various neuromuscular disorders**

Peter Van den Bergh, Belgium, discussed the development of the first Rasch-built measures addressing activity limitations (ACTIVLIM) in both adults and children with various neuromuscular disorders in order to follow patients' evolution over time. This measurement fulfilled basic and Rasch model requirements.<sup>13</sup> In addition, their experience through the years using this outcome measure and efforts to introduce its use nationally in the 6 accredited neuromuscular reference centres by means of a dedicated Belgian neuromuscular Disease Registry website application were addressed. Also, the availability of obtaining an online score for the ACTIVLIM was shown. It was agreed upon, that the item weights obtained with this scale differed from those obtained from the GBS, CIDP, and MGUSP specific R-ODS.<sup>2</sup> The differences were explained by the differences in cohorts of patients examined (ACTIVLIM examining patients with all forms of neuromuscular illnesses, R-ODS being disease specific). Finally, the construction of the ABILHAND measure through Rasch analyses, a tool to assess upper limb activity limitations was also discussed.<sup>14</sup>

### **The concept of MCID and Defining a responder**

Ingemar Merkies, the Netherlands, explained the concept of MCID and defining a responder. Traditional clinical trials tend to demonstrate the efficacy of a particular intervention by looking at p-values for groups' comparison or by comparing proportions of patients reaching an arbitrary pre-defined meaningful cut-off for an endpoint of interest. However, all these comparisons assume a fixed standard error (SE) across the metric being used. The concept of MCID was also addressed, highlighting its complexity due to the various available techniques.<sup>15</sup> The MCID represents a change that is considered meaningful and worthwhile by the patient such that they would consider repeating the intervention.<sup>16</sup> The MCID anchor-based methods and distribution-based methods were discussed.<sup>15,17</sup> Currently, there is no consensus regarding which MCID technique should be used.

Responsiveness techniques like the effect size can be misleading when doing groups comparison of the calculated responsiveness scores found between outcome measures in the same population at study.<sup>18</sup> However, using modern techniques like the Rasch method, comparison at the 'individual level of being a responder' can be made by using the model's provided individual standard errors for each patient at the various assessment time points. Outcome measures evaluated through Rasch demonstrate a 'U'-shape SE across the outcome measure range being used (see Figure 2 for an example).<sup>18,19</sup>

For the comparison purposes of the PeriNomS study, a responder was defined as a patient in whom the clinical condition improved enough to fall outside the MCID threshold boundaries of  $1.96 \times SE$  for the scale at study (see Figure 2C).<sup>18,19</sup>

By using this method, responders could be determined at the individual level in

addition to groups' comparison (Figure 3). For the Jamar versus Vigorimeter comparison, a combined (1 anchored-based and 1 distribution-based MCID technique) approach was used as previously suggested.<sup>17</sup>

## **Impairment measures**

### **MRC grading system: Original versus Rasch-built modification**

Ingemar Merkies, the Netherlands, presented data on the ordinal-based 6-point MRC grading system, showing that the MRC did not meet Rasch model's expectations, despite being used over 7 decades and in various neuromuscular clinical trials. In 1065 patients, with 7 different neuromuscular illnesses, almost 80% of all muscles were not properly scored by physicians, most showing disordered thresholds. The ability to properly discriminate between the different response options was hardly influenced by factors like experience of the physician and type of illness. Using the Rasch model, the 6-point MRC score was modified to a proposed four-points grading system.<sup>20</sup> Its use is, however, suggested only as part of Rasch-transformed composite scores in future clinical studies, since the four response categories are still ordinal based.

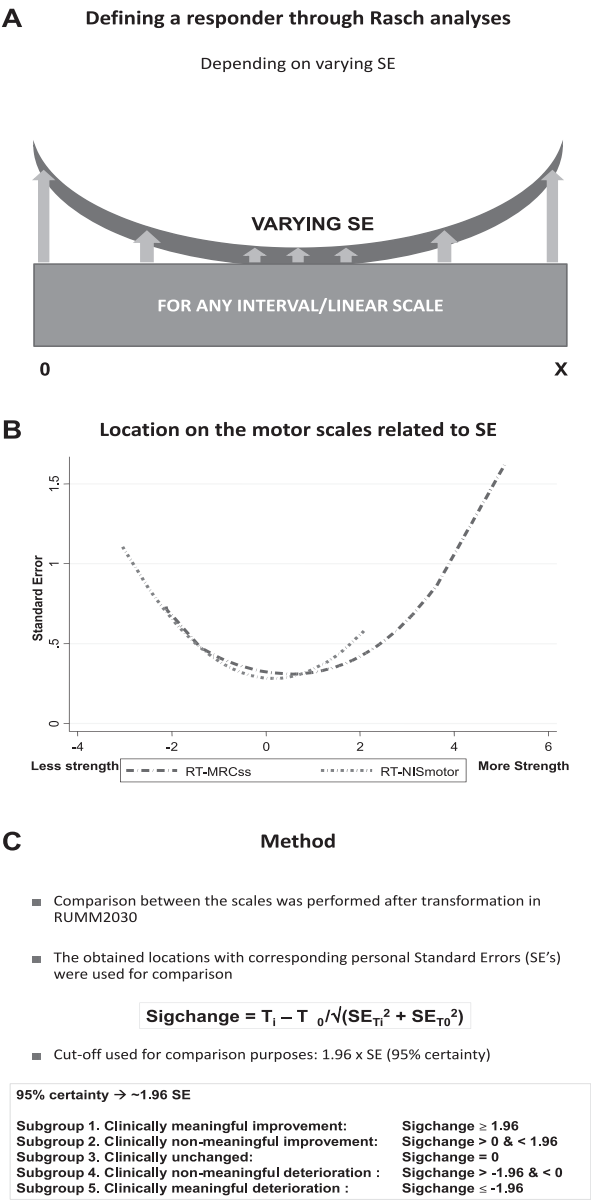
### **MRC sum score versus NIS motor subset**

Els Vanhoutte, the Netherlands, presented the equivalent validity and reliability findings in the cross-sectional comparison of the NIS motor subset versus the MRC sum score. Comparison between the scales was performed after Rasch transformation (RT-) and modifying the data as much as possible to meet model requirements, as important steps preceding comparison of the findings. The significant meaningful improvement (SigChange) was equivalent between the RT-MRC sumscore and the RT-NIS motor subset. SigChange for motor scales was mainly seen in patients with GBS (50%) and to a lesser extent in CIDP (20%). For both scales, SigChange was hardly seen in MMN. There was no SigChange found in MGUSP, but most of these patients received no active therapy.

### **Modified INCAT sensory scale versus NIS sensory subset**

Els Vanhoutte, the Netherlands, presented the comparison data between the modified ISS versus the NIS sensory subset. The ISS was modified (mISS) by incorporating in a standardized manner the examination of light touch sense, joint position sense, and the 2-point discrimination cut-offs that were based on recruited normative values.<sup>21</sup> Differences in their construction were addressed (mISS having a proximal gradient in its assessment versus sensory examination being only distally located in the NIS). The NIS sensory subset and the mISS had equivalent validity and reliability findings with a slightly better targeting of the mISS (better fitting to the patients' location on the metric by the mISS items' thresholds).

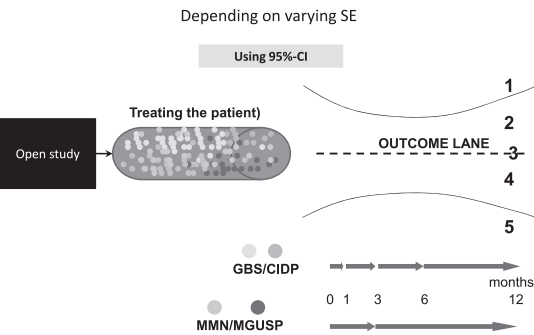
**Figure 2.** Demonstrating the U-shape SE pattern seen across interval/linear outcome measures



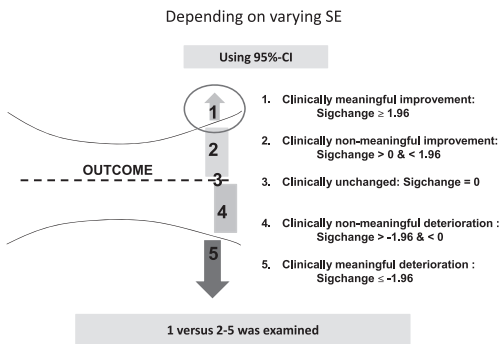
**Legend to Figure 2:** A: Schematic representation of the U-shape SE pattern across an outcome measure of interest; B: SE findings in two Rasch-Transformed motor composite scales; C: formula used to define meaningful improvement being defined as ≥ 1.96X SE for each individual patient, at a chosen of 95% confidence interval.<sup>18,19</sup> RT = Rasch transformed; MRCss = Medical Research Council sum score; NIS = Neuropathy Impairment Scale; SE = standard error.

**Figure 3.** Explaining the use of the individually obtained SE through Rasch to determine ‘being a responder’

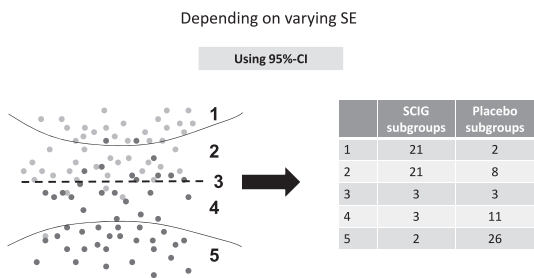
**A Defining a responder through Rasch analyses**



**B Defining a responder through Rasch analyses**



**C Defining a responder through Rasch analyses**



**Legend to Figure 3:** From the formula described in Figure 2C, five subgroups of patients can be determined; a cut-off of  $1.96 \times SE$  was applied for all comparison studies as part of the PeriNomS workshop. Since we were seeking for improvement in the PeriNomS study, all individuals exceeding  $+1.96 \times SE$  were considered having a meaningful improvement;  $18,19 SE =$  standard error. The figure should be read from A  $\rightarrow$  B  $\rightarrow$  C.



Thereafter, Kenneth Gorson, US, showed that the proportion of patients reaching the significant change cut-off ( $\leq -1.96 \times \text{SE}$ ; SigChange) were higher for the mISS compared to the NIS sensory subset. Furthermore, SigChange was mainly seen in patients with GBS (40%) (in CIDP only 10%). The SigChange was even less in patients with MGUSP (<10%).

### **Jamar versus Vigorimeter**

Els Vanhoutte, the Netherlands, showed that there were no differences in validity and reliability aspects between the Martin Vigorimeter and the Jamar. However, in all illnesses, most patients preferred the Vigorimeter. Since both instruments are at the ratio level, a combined MCID anchor-based (SF-36, question number 2) and a MCID distribution-based approach (unified rule of  $\frac{1}{2} \times \text{SD}$ ) were adopted for comparison purposes.<sup>17</sup> In essence, a patient was labelled as being a responder, if both MCID cut-offs were exceeded. Significant meaningful grip strength changes were equivalent between the two instruments. The longitudinal findings validated the suggested “many faces” of MCID, since the magnitude of the results varied depending on the MCID technique used.<sup>15</sup> Meaningful changes were mainly seen in patients with GBS (up to 81% with the Vigorimeter and 84% with the Jamar at 1 year follow-up), followed by the patients with CIDP (~40% for both instruments), SigChange was relatively low in MMN (Vigorimeter: 23.5%, Jamar 12.5% at 1 year follow-up) and hardly seen in MGUSP (6.7% versus 13.3%, respectively). The preference of patients for the Martin Vigorimeter was validated in the longitudinal group of patients throughout all illnesses.

### **Is there a place for skin biopsy as an outcome in inflammatory neuropathies?**

Giuseppe Lauria, Italy, discussed the state of the art in skin biopsies, focusing on facts and pitfalls. Skin biopsy for determining intraepidermal nerve fiber (IENF) density may be used in follow-up studies, as has been shown in some experimental studies. For clinical use, age and gender-adjusted normative data for the IENF density at the distal leg are available.<sup>22</sup> IENF density at the distal leg was reduced in GBS, CIDP and anti-MAG inflammatory neuropathies.<sup>23</sup> Furthermore, the IENF density at the distal leg correlated with early pain in GBS.<sup>23</sup> Several questions remain to be answered such as right-to-left variability, intra- and inter-individual variability over time and normal values for IENF density at proximal sites.

Dermal nerves may also reveal important information,<sup>24</sup> particularly in inflammatory neuropathies like GBS, CIDP, and MGUSP. However, the value of morphological changes, length measurement, myelin disruption, Ig, cell and complement depositions in the skin need further validation. In addition, the recruitment of normative values for dermal nerve assessment and determining potential differences in counting between laboratories were discussed. A place for skin biopsy findings as secondary outcome or exploratory outcome was proposed.

### **Is there a place for EMG as an outcome in inflammatory neuropathies?**

Vera Bril, Canada, discussed nerve conduction studies (NCS) in inflammatory neuropathies mainly focusing on the results of the ICE trial.<sup>25</sup> In addition, the value of having a central core lab for quality control and guidance purposes of NCS in multi-centre trials was discussed. NCS are necessary for diagnosis, and are considered objective measures that reflect clinical change and provide supportive evidence for positive therapeutic results. The use of NCS as an endpoint is also proposed by the FDA. NCS as a secondary outcome measure was proposed.

### **Pathophysiology of CIDP**

Ivo van Schaik, the Netherlands, questioned whether we should re-focus on unravelling pathophysiological mechanisms underpinning inflammatory neuropathies. A historical overview of pathophysiological studies including possible biomarkers in inflammatory neuropathies was discussed. Understanding all these may contribute to determine future trial strategies and/or directions for treatment regimens.

Biomarkers could be useful to define subgroups with different pathogenic mechanisms and with different responses to treatment. Different possible antibodies and genetic factors were discussed, particularly highlighting the approach shown recently in a group of patients with anti-contactin antibodies.<sup>26</sup> The need for CIDP (and other inflammatory neuropathies) registry and biobanking was stated.

## **Activity and participation**

### **Disability Measures in inflammatory neuropathy trials**

Michael Lunn, UK, discussed the definition of disability as presented by the WHO in 1980 and 2001 and then gave an overview of the disability measures used in inflammatory neuropathy trials up to 2013.<sup>10</sup> The term disability was replaced by the positive concept of 'activities' in the 2001 WHO definitions and is described as 'the nature and extent of executing a task or action by an individual'. The clinimetric aspects of the disability measures used in trials so far were discussed and the advantages and disadvantages of classical test theory and Rasch built scales were highlighted. In turn, the GBS Disability scores, the modified Rankin score, the functional Independence measure, the Rivermead Mobility Index, the Guy's originated disability scores (GNDS, INCAT, ODSS and ONLS), the AMC linear disability scale,<sup>27</sup> the ALDS item bank,<sup>28</sup> the ACTIVLIM<sup>13</sup> and the R-ODS<sup>2)</sup> scales were illustrated in detail. Lunn concluded that activity and participation should be the primary outcome level in trials in inflammatory neuropathies mainly because of its meaning to patients and care-givers, and the general ease and applicability of scales in this domain.

### **The R-ODS family: their construct and (dis)advantages.**

Els Vanhoutte, the Netherlands, presented the development of a patient-based, linearly-weighted Rasch-built overall disability scale (R-ODS) and the evaluation of its clinimetric properties. The 24-item R-ODS for GBS, CIDP and MGUSP fulfilled all Rasch model expectations.<sup>2</sup> However, the R-ODS as it is currently used showed item bias in 6 of the 24 (25%) items when examining cultural differences. Cross-cultural validation is mandatory for international use of any scale. Therefore, re-modelling of the R-ODS should be undertaken. The most widely used activity scale in GBS and CIDP is the INCAT 10-points disability measure.<sup>29</sup> Its construct is quite similar to the ODSS and ONLS. For comparison purposes, the collected longitudinal data of the ONLS was transformed through Rasch and demonstrated improper targeting with large gaps between items, local dependency, and misfit statistics for all items. Comparison between the RT-ONLS and the R-ODS showed a significant higher proportion of meaningful improvement (SigChange; cut-off  $\geq 1.96 \times \text{SE}$ ) in the R-ODS when compared to INCAT-ODSS-ONLS. With respect to the R-ODS, a higher proportion of SigChange was seen in patients with GBS (up to 100% at 1 year of follow-up; INCAT-ONLS: 80%) compared to CIDP (~45% at 1 year follow-up; INCAT-ONLS 28%). The SigChange was poor in patients with MGUSP (at 3 months 20%; most probably determined by the few patients that received therapy; at 1 year 6%).

### **The R-ODS family – for MMN**

Ingemar Merkies, the Netherlands, demonstrated the composition of the R-ODS specifically for MMN including its validation, reliability and responsiveness findings using Rasch. A relatively low SigChange was seen in the R-ODS for MMN (only 20% being a responder at 1 year; cut-off  $1.96 \times \text{SE}$ ). This was most probably explained by most patients not being naive treated patients. Based on these findings the following suggestions were made for evaluating patients with MMN:

- Perhaps examining 'being a responder' should be performed using a much lower cut-off, e.g. at  $1 \times \text{SE}$  or the definition of being a responder should be revisited, not only including the subgroup of significant improvement, but also non-significant improvement, and stability (i.e. responders are those patients that do not deteriorate).
- MMN (and MGUSP) should be examined for a longer period over time due to their indolent course to capture relevant changes.

The group discussed that in naive MMN patients significant improvement should be sought for. In addition, in those patients who are receiving maintenance therapy, being a responder could be applied in the broader definition as presented above. A self-evaluation scale (SES) in MMN, based on the Canadian occupational Performance Measure (COPM)<sup>30</sup> demonstrated ~20% SigChange (1 year follow-up; cut-off  $1.96 \times \text{SE}$ ). However, since the items vary per patient, no adaptations to Rasch model could be achieved using this outcome measure.

## Quality of Life

### Patients perspective

Glennys Sanders, UK, urged for the recognition of the importance of patients' experiences, expectations and emotions. Physicians should take factors like motivation, fatigue, pain and acceptance into account, since these factors can cause depression and affect concentration and confidence to focus on the challenges of their recovery. Furthermore these factors contribute significantly to disability and quality of life reduction. The importance of a positive and supportive/encouraging approach towards patients with inflammatory neuropathies was stressed. Attention for the impact of those receiving respiratory ventilation as well as for those being not able to smile were addressed. The group acknowledged that more attention for these aspects should be the focus of future studies in these conditions. Patients need an ability to express their anger, frustration and fears both for the present and future. Therefore, Sanders emphasised that all patients should be guided to relevant Support Groups to enable them to share knowledge, experiences and emotions.

### Recommendations

After taking all the presented results into account, the group aimed to get consensus on a specific core set of outcome measures to be used in all future trials in patients with GBS, CIDP, MGUSP, and MMN separately.

Table 1 provides an overview of the recommendations.

#### GBS

The primary outcome in GBS in future trials should be at the activity and participation level measured by the inflammatory R-ODS,<sup>2</sup> since it has demonstrated significantly higher responsiveness compared to the INCAT-ONLS.

- **Minimal core set for future GBS studies**

The minimum core set should include:

- At the impairment level: the Martin Vigorimeter, the Rasch-transformed-mISS, being ventilated (Y/N) and the duration of respiratory ventilation.
- At the activity and participation level: the R-ODS and the GBS disability scale. The latter was suggested for historical purposes.

- **Recommendations in GBS**

It was recommended that future trials should also include the RT-FSS,<sup>31</sup> the RT-MRC sum score, and the original MRC sum score, although no consensus was reached on this.

- **Future needs in GBS**

Pain should be further evaluated in terms of its underlying pathophysiological

mechanism as well as in terms of evaluating proper therapy. Also, measuring strength needs further exploration, since there was no consensus how, and if, strength should be measured; the possibility of including a muscle dynamometer and the RT-MRC sum score were discussed. The R-ODS could be further improved, by cross-cultural modification. At the quality of life level, a Rasch-transformed outcome measure based on the future findings of the PeriNomS study should be included.

## CIDP

The primary outcome in CIDP in future trials should be at the activity and participation level measured by the R-ODS,<sup>2</sup> since it has demonstrated significantly higher responsiveness compared to the INCAT-ONLS.

- **Minimum core set for future CIDP studies**

The minimum core set should include:

- At the impairment level, the Martin Vigorimeter, the RT-mISS, and a 'manual muscle testing' procedure, not otherwise defined.
- At the activity and participation level: the R-ODS and the original INCAT disability score.
- At the quality of life level, the 5-points patient global impression of change (PGIC)<sup>32</sup> and SF-36<sup>33</sup> should be used until the data of the PeriNomS study regarding quality of life measures are further analysed.

- **Recommendations in CIDP**

It was recommended to use the RT-FSS and to include pain measurements such as the 11-point Pain-Intensity Numerical Rating Scale (PI-NRS).<sup>32</sup>

- **Future needs in CIDP**

The RT-MRC sum score needs further development, and should possibly include additional muscles. The criteria for measuring pain should be examined. A walking test (2 or 6 minutes) could be considered. The R-ODS could be further improved, by cross-cultural modification. At the quality of life level, a Rasch-transformed outcome measure based on the future findings of the PeriNomS study should be included.

## MMN

The primary outcome in MMN in future trials should be at the activity and participation level measured by the disease specific R-ODS.<sup>34</sup>

- **Minimum core set for future MMN studies**

- At the impairment level, the Vigorimeter should be used. In addition, patient-specific affected muscles should be incorporated and future studies are needed to re-determine these patient-specific MRC sum scores that may vary from patient to patient. The use of RT-MRC scores was proposed.
- At the activity and participation level, the RODS-MMN was suggested.
- At the quality of life level, a Rasch-transformed outcome measure based on the future findings of the PeriNomS study.

- **Future needs in MMN**

The R-ODS could be expanded, possibly using the ALDS<sup>27</sup> or the ABILHAND.<sup>14</sup> At the quality of life level, a Rasch-transformed outcome measure based on the future findings of the PeriNomS study should be included.

## **MGUSP**

- **Recommended outcome measures for future MGUSP studies**

Changes were hardly captured in MGUSP patients. Since most patients have not received any therapy, the obtained data could be considered as being historical. Therefore, a *core set* could not be defined. The group agreed upon extending the evaluation of patients with MGUSP over a much longer period. In addition, more data on a larger sample size receiving therapy is needed before a minimum core set of outcome measures for future trials can be defined.

However, at this point, the RT-mISS was recommended at the impairment level, and the R-ODS was recommended at the activity and participation level (the latter as primary outcome suggested). Also the original INCAT 10-point disability should be included based on significant p-values seen in two previous trials<sup>35</sup> (1 not published yet). At the quality of life level, the PGIC and the SF-36 or Euro-QoL were arbitrarily recommended.

- **Future needs in MGUSP**

Extending the evaluation of patients with MGUSP over a longer period is needed. More treated patients should also be examined. At the impairment level, pain evaluation, ataxia and a tremor score as well as the 9-hole PEG test could be further explored. At the quality of life level, a Rasch-transformed outcome measure based on the future findings of the PeriNomS study should be included.

## **Overall and Closure**

Consensus was reached for most illnesses at the various levels of assessing outcome, except regarding whether there is a need to perform manual muscle testing and how to handle the data. The Rasch-built activity and participation measures were proposed as the future primary outcome measures in all illnesses. Recommendations were made for additional evaluation of aspects like fatigue, pain, and mobility, and future needs were systematically highlighted. Additional evaluation in MGUSP over a longer period of time, particularly in patients receiving therapy was also agreed upon. The PeriNomS group will also be focusing on determining the impact of collected skin biopsy data as well as the construction of an inflammatory neuropathy specific Rasch-built quality of life metric.

Last but not least, the PeriNomS study is performing more than collecting and evaluating data in inflammatory neuropathies. The PeriNomS contributes to the

wrinkles needed to create a paradigm shift in outcome assessment in future peripheral neuropathic studies and could serve as an example for other illnesses in neurology and medicine in general.

**Table 1.** Overview of the minimum core set, recommendations, and future needs

	GBS	CIDP	MMN	MGUSP
<b>Minimal core set</b>				
Impairment level	Martin Vigorimeter RT-mISS Being ventilated (Y/N) Duration of ventilation	Martin Vigorimeter RT-mISS 'Manual muscle testing'	Martin Vigorimeter Patient-specific muscle testing RT-MRC scores	not yet defined; further study required
Activity and participation level	R-ODS* GBS disability scale	R-ODS* Original INCAT disability score	R-ODS MMN*	see above
Quality of life level		5-PGIC SF-36	RT-QoL scale	see above
<b>Recommendations</b>				
Impairment level	RT-MRCss Original MRCss RT-FSS	11-PI-NRS RT-FSS	-	RT-mISS
Activity and participation level		-	-	R-ODS Original INCAT 10-point
Quality of life level		-	-	PGIC SF36 or Euro-QoL
<b>Future needs</b>				
Impairment level	Pain Muscle dynamometer/ RT-MRCss	RT-MRCss Pain Walking test	-	Define core set Pain Ataxia Tremor 9-hole PEG test
Activity and participation level	Cross-cultural R-ODS	Cross-cultural R-ODS	expanding the R-ODS	Define core set
Quality of life level	RT-QoL scale	RT-QoL scale	RT-QoL scale	RT-QoL scale

**Legend to Table 1:** RT = Rasch transformed; mISS = modified INCAT sensory sumscore; R-ODS = Rasch-built overall disability scale; MRCss = Medical Research Council sum score; FSS = Fatigue Severity Scale; 5-PGIC = 5-points patient global impression of change; 11-PI-NRS = 11-point Pain-Intensity Numerical Rating Scale; QoL = Quality of Life.

## References

- 1 Merkies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 2 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 3 van Nes, S. I., Faber, C. G., and Merkies, I. S. Outcome measures in immune-mediated neuropathies: the need to standardize their use and to understand the clinimetric essentials. *J Peripher Nerv Syst* **13**, 136-147 (2008).
- 4 Dyck, P. J., Boes, C. J., Mulder, D. *et al.* History of standard scoring, notation, and summation of neuromuscular signs. A current survey and recommendation. *J Peripher Nerv Syst* **10**, 158-173 (2005).
- 5 Kleyweg, R. P., van der Meche, F. G., and Schmitz, P. I. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle Nerve* **14**, 1103-1109 (1991).
- 6 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 7 Streiner, D. L. and Norman, G. R. *Health measurement scales: a practical guide to their development and use*. 2nd edn, (Oxford University Press, USA, 1998).
- 8 Latov, N., Deng, C., Dalakas, M. C. *et al.* Timing and course of clinical response to intravenous immunoglobulin in chronic inflammatory demyelinating polyradiculoneuropathy. *Arch Neurol* **67**, 802-807 (2010).
- 9 DeVellis, R. F. Classical test theory. *Med Care* **44**, S50-59 (2006).
- 10 World Health Organisation. The International Classification of Functioning, Disability and Health (ICF). (WHO, 2001).
- 11 Tennant, A. and Conaghan, P. G. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* **57**, 1358-1362 (2007).
- 12 Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (University of Chicago Press, 1960).
- 13 Vandervelde, L., Van den Bergh, P. Y., Goemans, N. *et al.* ACTIVLIM: a Rasch-built measure of activity limitations in children and adults with neuromuscular disorders. *Neuromuscul Disord* **17**, 459-469 (2007).
- 14 Wang, T. N., Lin, K. C., Wu, C. Y. *et al.* Validity, responsiveness, and clinically important difference of the ABILHAND questionnaire in patients with stroke. *Arch Phys Med Rehabil* **92**, 1086-1091 (2011).
- 15 Beaton, D. E., Boers, M., and Wells, G. A. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol* **14**, 109-114 (2002).
- 16 Jaeschke, R., Singer, J., and Guyatt, G. H. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* **10**, 407-415 (1989).
- 17 Merkies, I. S., van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *J Neurol Neurosurg Psychiatry* **81**, 1194-1199 (2010).
- 18 Hobart, J. and Cano, S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* **13**, iii, ix-x, 1-177 (2009).
- 19 Wright, B. D. and Stone, M. H. *Best test design*. (Mesa Press, 1979).
- 20 Vanhoutte, E. K., Faber, C. G., van Nes, S. I. *et al.* Modifying the Medical Research Council grading system through Rasch analyses. *Brain* **135**, 1639-1649 (2012).
- 21 van Nes, S. I., Faber, C. G., Hamers, R. M. *et al.* Revising two-point discrimination assessment in normal aging and in patients with polyneuropathies. *J Neurol Neurosurg Psychiatry* **79**, 832-834 (2008).
- 22 Lauria, G., Bakkers, M., Schmitz, C. *et al.* Intraepidermal nerve fiber density at the distal leg: a worldwide normative reference study. *J Peripher Nerv Syst* **15**, 202-207 (2010).
- 23 Ruts, L., van Doorn, P. A., Lombardi, R. *et al.* Unmyelinated and myelinated skin nerve damage in Guillain-Barre syndrome: correlation with pain and recovery. *Pain* **153**, 399-409 (2012).



- 24 Lauria, G., Cazzato, D., Porretta-Serapiglia, C. *et al.* Morphometry of dermal nerve fibers in human skin. *Neurology* **77**, 242-249 (2011).
- 25 Bril, V., Banach, M., Dalakas, M. C. *et al.* Electrophysiologic correlations with clinical outcomes in CIDP. *Muscle Nerve* **42**, 492-497 (2010).
- 26 Querol, L., Nogales-Gadea, G., Rojas-Garcia, R. *et al.* Antibodies to contactin-1 in chronic inflammatory demyelinating polyneuropathy. *Annals of neurology* **73**, 370-380 (2013).
- 27 Holman, R., Lindeboom, R., Vermeulen, M. *et al.* The AMC Linear Disability Score project in a population requiring residential care: psychometric properties. *Health Qual Life Outcomes* **2**, 42 (2004).
- 28 Holman, R., Weisscher, N., Glas, C. A. *et al.* The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. *Health Qual Life Outcomes* **3**, 83 (2005).
- 29 Hughes, R., Bensa, S., Willison, H. *et al.* Randomized controlled trial of intravenous immunoglobulin versus oral prednisolone in chronic inflammatory demyelinating polyradiculoneuropathy. *Annals of neurology* **50**, 195-201 (2001).
- 30 Leger, J. M., Chassande, B., Musset, L. *et al.* Intravenous immunoglobulin therapy in multifocal motor neuropathy: a double-blind, placebo-controlled study. *Brain* **124**, 145-153 (2001).
- 31 van Nes, S. I., Vanhoutte, E. K., Faber, C. G. *et al.* Improving fatigue assessment in immune-mediated neuropathies: the modified Rasch-built fatigue severity scale. *J Peripher Nerv Syst* **14**, 268-278 (2009).
- 32 Farrar, J. T., Young, J. P., Jr., LaMoreaux, L. *et al.* Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* **94**, 149-158 (2001).
- 33 Ware, J. E., Jr., Kosinski, M., and Gandek, B. *SF-36® Health Survey. Manual and interpretation guide.*, (QualityMetric Incorporated, 2000).
- 34 Vanhoutte, E. K., van Nes, S. I., Cats, E. A. *et al.* Rasch-Built overall disability scale for multifocal motor neuropathy (RODS-MMN). Abstracts of the 2011 Meeting of the Peripheral Nerve Society. June 25-29, 2011. Potomac, Maryland, USA. *J Peripher Nerv Syst* **16 Suppl 3**, S1-174 (2011).
- 35 Dalakas, M. C., Rakocevic, G., Salajegheh, M. *et al.* Placebo-controlled trial of rituximab in IgM anti-myelin-associated glycoprotein antibody demyelinating neuropathy. *Annals of neurology* **65**, 286-293 (2009).

# **Summary, discussion and future perspectives**

## General introduction and background information

This thesis presents a series of studies on outcome measures in immune-mediated neuropathies (Guillain-Barré syndrome (GBS), chronic inflammatory demyelination polyradiculoneuropathy (CIDP), IgM monoclonal gammopathy of unknown significance polyneuropathy (MGUSP) and multifocal motor neuropathy (MMN). The aim of the Peripheral Neuropathy Outcome Measures Standardization (PeriNomS) study (**Chapter 1**) is to present a minimum core set of outcome measures to be used in future clinical trials and follow-up studies in patients with immune-mediated polyneuropathy. The PeriNomS study embraces three parts. The first part was conducted by dr. Sonja I. van Nes who defended her thesis in January 2011, addressing normative value studies, comparative validity and reliability studies, creating several outcome measures based on the Rasch model and introducing the concept of the 'minimum clinically important difference' (MCID).<sup>1-6</sup> This thesis presents the results of the second part of the PeriNomS study: 163 patients were serially examined over a one-year period. The primary aim was to establish through comparative responsiveness of the various outcome measures a core set of outcome measures in these disorders that should be used in all future trials in these conditions as part of a consensus. Finally, the remaining future needs as the 3<sup>rd</sup> part of the PeriNomS will be addressed in "future perspectives".

The design of PeriNomS study is also described in **Chapter 1**. For its cross-sectional and longitudinal parts, several outcome measures were examined at different levels of assessing outcome according to the WHO international classification and quality of life concept (pathology, impairment, activity limitation and quality of life) taking into account the input by various experts in the field.<sup>7,8</sup> Furthermore, the clinical characteristics of the immune-mediated polyneuropathies, GBS, CIDP, MGUSP and MMN are briefly described.

Most outcome measures used thus far in inflammatory neuropathies are at the ordinal level and based on the classical test theory, despite their known deficiencies. The Rasch model overcomes the disadvantages of ordinal-based outcome measures (**Chapter 2**). This theory is based on a logical assumption: a patient with a high ability will have a higher probability of affirming a particular task compared to a patient with a lower ability. Various check-up points (e.g. good item and person statistical fit, threshold ordering, no item bias or local dependency, and demonstrating unidimensionality) in the construction and evaluation of outcome measures using the Rasch method are highlighted. When these Rasch model's expectations are fulfilled, measurement at the interval level is obtained.

Trial results evaluating treatment options are often driven by the presence or absence of having a significant (p-value) difference between the various (e.g. treated versus placebo) groups. However, the presence of a statistical significant difference does

not always mean that the findings are clinically relevant. The concept of minimum clinically important difference (MCID) is introduced in **Chapter 3** and is defined as the minimum change score necessary to reflect a clinically relevant change.<sup>9</sup> There are several methods available to calculate the MCID, thus showing its many faces and the lack of international consensus on this matter.<sup>10</sup> The biggest disadvantages of applying these MCID techniques is perhaps the standard error (SE) being considered “static (unchanged)” throughout the range of the outcome measure being used. The MCID-SE concept using Rasch-transformed data demonstrates a dynamic pattern of ‘being-a-responder’ in patients with immune-mediated neuropathies. SE values varied throughout a metric’s continuum and are the lowest in the middle part and increases towards the edges of the metric’s range. A responder is defined if a patient showed a significant change exceeding  $\text{MCID-SE} \geq 1.96$ .

## Impairment studies

Based on the WHO, impairment is a problem in body function or structure that may contribute to activity and participation restrictions.<sup>8</sup> Fatigue, an impairment quality, is considered a major disabling complaint in patients with immune-mediated neuropathies.<sup>11</sup> The 9-item fatigue severity scale (FSS) has been introduced to assess the severity of fatigue and has demonstrated its scientific soundness in immune-mediated neuropathies. The FSS is an ordinal outcome measure based on the classical test theory, therefore it is not surprising that the original 9-item FSS did not meet the Rasch model expectations (**Chapter 4**). Data from the INCAT studies in patients with GBS, CIDP, and MGUSP were subjected to Rasch analyses. After removing two misfitting items and collapsing the 7 response categories to 4, a more homogenous 7-item Rasch-built fatigue scale was constructed that fulfilled all Rasch model’s requirements. Since the Rasch-Transformed FSS met the Rasch-Rating scale model’s requirements, its use is recommended in future clinical trials when using sumscores to compare e.g. the effect between treatment and placebo on fatigue.

Muscle strength and sensory measurement are frequently used in clinical trials as a primary outcome measure. The modified ISS (mISS) and the Neuropathy Impairment Score sensory subset (NISs) are commonly used sensory measurements. Muscle strength is often measured using a tool like the Martin Vigorimeter or by manual muscle testing (Medical Research Council (MRC) grades or Neuropathy Impairment Score motor subset (NISm)).<sup>12-19</sup> The sensitivity of assessing grip strength with the Martin Vigorimeter is demonstrated using the data obtained from the ICE trial, the largest randomized trial ever performed in patients with CIDP. The Vigorimeter demonstrated the ability to capture clinically meaningful changes quite early in the interventional phase when compared with the primary ordinal-based outcome measure used

(**Chapter 5**).<sup>20</sup> The MRC is the most widely applied tool to manually measure strength of muscle groups at bedside.<sup>21</sup> In **Chapter 6** it is demonstrated that physicians, independent of their experience or the type of neuromuscular illness examined, are unable to apply the MRC grades in a proper manner. In 86% of the examined muscle groups in 1065 patients we found disordered thresholds.<sup>22</sup> Also, the MRC grading system failed to differentiate between various degrees of muscle weakness. In order to solve these problems, the MRC grades were modified to a homogenous four category response options. This might intuitively lower the ability to capture minor functional changes in a patient. However, keeping the (disordered) six responses will give a false sense of precision and potentially increase the error in assessment, which may lead to a false sense of clinically meaningful improvement. The four response categories are still ordinal based; therefore, meaningful sum scores can only be made after transforming the data to the interval level through Rasch method. After Rasch modelling, we were able to present a transformed modified MRC 12-muscle group summed score for GBS and CIDP.

The NISs versus the mISS and the NISm versus the MRC sum score were compared after being subjected to Rasch (**Chapter 7**). Longitudinally obtained data in the transformed scales were also compared with each other in terms of responsiveness using the concept of MCID-SE. Responsiveness was poor in the sensory measures and moderate in the motor scales. Motor scales demonstrated higher values in GBS compared to CIDP. Overall, no differences were seen when comparing the sensory scales with each other or when comparing the motor scales' findings. In patients with MGUSP, motor and sensory responsiveness, if any, were very poor.

## Activity and participation studies

Based on the World Health Organisation definition, activity is the execution of a task or action by an individual, while participation is involvement in life situations.<sup>8</sup> Since there is no clear-cut boundary between activities and participations, there should be one single outcome measure at the activity/participation level that can be used in all immune-mediated neuropathies.

Though GBS/CIDP and MGUSP are different disorders, especially in terms of natural course, Rasch analysis showed that there is no DIF, meaning that patients with these disorders judge items the same way. Therefore, the RODS for GBS, CIDP and MGUSP is the same. MGUSP is far more indolent, which is reflected in a lower responsiveness. However, patients with MMN experience items differently, therefore necessitating a different outcome scale. Though this may seem a disadvantage, anchoring may provide the opportunity to compare the outcomes between these disorders.

Therefore, two different Rasch-built overall disability scales (RODS) were developed that fulfilled the Rasch model's requirements: one for GBS/CIDP/MGUS (inflammatory-

RODS) (**Chapter 8**) and one specifically for MMN (**Chapter 9**) (MMN-RODS). In these chapters, the steps taken to their construction are described, also highlighting some methodological issues found. Eventually, both measures fulfilled all model's expectations, being the first interval scales specifically illness-designed at the activity and participation level in these conditions.

Compared to the Overall Disability Sum Score, the Inflammatory-RODS (I-RODS) represented a wider range of item difficulties, thus showing better targeting of patients with different ability levels. In addition, less ceiling and floor effects were seen with the I-RODS.<sup>5</sup> However, cross-cultural examination of the 24 items revealed item bias in 5 of the 24 items. This emphasizes the importance of cross-cultural validation before applying an outcome measure in different geographical regions. Despite these findings, the longitudinal international part of the PeriNomS study demonstrated more often clinically meaningful changes over time in the I-RODS, with a greater magnitude of change, when compared to the INCAT-ONLS disability scale in patients with GBS and CIDP. For IgM-MGUSP, the lack of responsiveness during the 1-year study did not allow a clear separation between the scales (**Chapter 10**). In MMN, the responsiveness of the MMN-RODS was relatively low (**Chapter 9**).

The MCID-SE concept using Rasch-transformed data demonstrated a dynamic pattern of 'being-a-responder' in patients with immune-mediated neuropathies. The I-RODS showed the highest responsiveness scores in patients with immune-mediated neuropathies compared to other general measures applied. The natural course of the disease needs to be taken into account when establishing a cut-off value for clinically important change. In disorders with an indolent nature like MGUSP, assessment may require a longer period or the development of more sensitive measures (**Chapter 11**).

## **The minimum core set of outcome measures for future studies in inflammatory neuropathies (Chapter 12)**

The participants (including a patients' representative) at the 196<sup>th</sup> ENMC International Workshop ('Outcome Measures in inflammatory peripheral Neuropathies') gathered during a 3 day meeting in 2013 and reached consensus on a specific core set of outcome measures to be used in all future trials in immune-mediated neuropathies as presented in Table 1.<sup>23</sup> The primary outcome in patients with GBS, CIDP and MMN should be at the activity and participation level measured with the constructed RODS illness-specific scales. Since changes were hardly captured in MGUSP patients, a definite core set could not be defined in this disorder. Patients with MGUSP should be evaluated over a longer period of time and more data on a larger sample size receiving therapy are needed to ultimately define a minimum core set of outcome measures in this disorder. The possibility of constructing an MGUSP specific scale was also addressed. At this point, in MGUSP patients the RT-mISS was recommended at the

impairment level, and the inflammatory-RODS was recommended at the activity and participation level (the latter as primary outcome suggested).

**Table 1.** Overview of the minimum core set and recommendations

	GBS	CIDP	MMN	MGUSP
<b>Minimal core set</b>				
Impairment level	Martin Vigorimeter RT-mISS Being ventilated (Y/N) Duration of ventilation	Martin Vigorimeter RT-mISS 'Manual muscle testing'	Martin Vigorimeter Patient-specific muscle testing RT-MRC scores	not yet defined; further study required
Activity and participation level	R-ODS* GBS disability scale	R-ODS* Original INCAT disability score	R-ODS MMN*	see above
Quality of life level		5-PGIC SF-36	RT-QoL scale	see above
<b>Recommendations</b>				
Impairment level	RT-MRCss Original MRCss RT-FSS	11-PI-NRS RT-FSS	-	RT-mISS
Activity and participation level		-	-	R-ODS Original INCAT 10-point
Quality of life level		-	-	PGIC SF36 or Euro-QoL

**Legend Table 1:** 5-PGIC = 5-points patient global impression of change; 11-PI-NRS = 11-point Pain-Intensity Numerical Rating Scale; FSS = Fatigue Severity Scale; INCAT= Inflammatory Neuropathy Cause and Treatment; I-RODS=inflammatory RODS; mISS = modified INCAT sensory sumscore; MRCss = Medical Research Council sum score; RODS = Rasch-built overall disability scale; RT = Rasch transformed; QoL = Quality of Life. \*Indicates the primary outcome measure.

## Future perspectives

### General

The choice of outcome measures in clinical trials may have an impact on the amount of responders that is captured. Interval measurement is preferred over ordinal-based scales since these are suitable for conventional statistics and allow for the creation of sum scores. One should be cautious when using sum scores based on ordinal scales. The ordering of the response options in a ordinal scale should be formally examined using Rasch before assuming that the sum of the parts equals the whole.<sup>24</sup> Using response options that are not working correctly may give a false sense of precision and could lead to conclusions based on a clinically meaningful improvement that may not exist.

Furthermore, a shift is needed from p-value to responsiveness at the *individual level* based on the MCID-SE using varying SE depending on the patients' ability. The cut-off

value for clinically important change should take the natural course of the disease into account.

MMN and MGUSP are indolent illnesses and change was hardly captured in a 1-year period. It might be necessary to consider lowering the cut-off value to 1xSE when examining a responder in these disorders. However, the research community should agree whether the higher degree of imprecise measurement followed by lowering the cut-off value to 1xSE is acceptable. Furthermore, physicians should consider increasing the follow-up period to a minimum of 2 years in these indolent disorders. In MMN, further research is needed to examine responsiveness at muscle level for the individual patient. In MGUSP there is a strong need to extend the evaluation period and to increase the number of patients included and to consider the development of a MGUSP specific RODS scale; this will be part of the PeriNomS study part III.

### **Outcome measures**

When designing clinical trials, researchers should strive for uniform outcome measures at the interval level in order to be able to compare future clinical trials and thereby improve the quality of randomised controlled trials in immune-mediated neuropathies.

- **Pathology**

- Intra-epidermal nerve fibers: Normative data for the intra-epidermal nerve fiber (IENF) density at the distal leg are available and currently used in clinical practice.<sup>25</sup> The IENF density might be a useful tool to predict pain and duration of complaints in patients with immune-mediated neuropathies. However, before the clinical applicability in immune-mediated neuropathies can be established, questions such as right-to-left variability, intra- and inter-individual variability over time and normal values for IENF density at proximal sites need to be answered.

- **Impairment**

- Muscle strength: Muscle testing could be a fast and easy applicable tool in daily clinical practice and trials. Currently, there is no consensus on how strength should be measured and if strength should be a primary outcome measure. It should be examined whether muscle testing is sensitive enough to capture improvement. A comparison with other methods (e.g. the RODS) should be made that might be more sensitive in capturing change. Possible strength measures include the use of a muscle dynamometer or the RT-MRC sum score. Although the Vigorimeter has been shown to be a prognostic indicator of clinical and functional recovery and is useful in monitoring the effect of treatment, it examines solely grip strength.<sup>26</sup> Therefore, the use of dynamometric devices that are able to assess impairment distally and proximally should be further explored. The RT-MRC sum score based on the proposed four-points grading system meets the Rasch-model requirements. However, the ability to capture clinically important change should be improved, especially in MMN. It has been suggested to examine



which muscles are most sensitive to capture change in each immune-mediated neuropathy. This might lead to the creation of a muscle-specific responsiveness scale. Longitudinal examination and responsiveness studies on muscle strength measurement should be performed.

- Nerve conduction studies (NCS) are necessary for diagnosis, and are considered objective measures that reflect clinical change and provide supportive evidence for positive therapeutic results. NCS as a secondary outcome measure was proposed but responsiveness studies are needed. In the ICE trial, various NCS parameters showed improvement in the treated arm, but the proportion of patients and magnitude of improvement in these parameters was relatively poor.<sup>27</sup>
- Pain is a factor that may contribute to activity and participation restrictions and a reduced quality of life. Therefore, it should be further evaluated in terms of its underlying pathophysiological mechanism as well as in terms of evaluating proper therapy. Also, the criteria for measuring pain should be examined. There are several methods available that can measure pain, e.g. the 11-point PI-NRS, the VAS or the PGIC. However, also the VAS and the PI-NRS are considered ordinal scales.<sup>28-30</sup> Further evaluation of these scales is needed and standardized studies to measure pain should be created. After this, the ability to capture changes in pain over time in patients with GBS, CIDP and MGUSP should be examined.
- Fatigue can cause severe activity and participation restrictions, but might not significantly contribute in the acute setting. The impact of fatigue in clinical trials is unknown and should be further explored. The 7-item Rasch-Built fatigue scale fulfilled all requirements of the Rasch model, however, responsiveness studies and cross-cultural validation should be performed.<sup>4</sup>
- **Activity and participation**
  - The different RODS scales have been suggested as the primary outcome measure in immune-mediated neuropathies.<sup>23</sup> However, development of a specific RODS for MGUSP could be considered, since hardly any change was seen in this disorder using the inflammatory-RODS. Furthermore, expanding the MMN-RODS using e.g. the ABILHAND might be considered.<sup>31,32</sup>
- **Quality of life:**
  - Analysing the quality-of-life (QoL) outcome measures is beyond the scope of this thesis and will be part of the PeriNomS study part III. A preliminary QoL scale that includes all questions of existing QoL outcome measures may be used to develop a Rasch-built QoL outcome measure specifically for inflammatory neuropathies.<sup>33-39</sup> Also, when creating the preliminary QoL questionnaire, one should take the patient's perspective into account.
- **Miscellaneous**
  - Intuitively, composite scales give a good reflection of what is important to the individual patient. However, since the items of the personal patient-centred

measures (PPCM) vary per patient, it is therefore logic to assume that fit to the Rasch models expectations will most probably not be achieved. The evaluation of the PPCM and the Clinical Judgment Score (ClinJSc) will be performed as part of the PeriNomS study part III.<sup>40,41</sup>

## Conclusion

Outcome measures should be taken seriously and proper outcome measures should be the basis for any future studies in patients with immune-mediated neuropathies, being preferably at the interval or ratio level. The Rasch method can be a tool to create new interval scales or transform an existing ordinal one into an interval scale. The use of uniform outcome measures is highly recommended in order to be able to compare future clinical trials and thereby improve the quality of randomised controlled trials in inflammatory neuropathies, enabling comparison between studies. The primary outcome measure in immune-mediated neuropathies should be at the activity and participation level using the disease-specific RODS questionnaires. The conclusion of clinical trials should be based on individual responsiveness (MCID-SE), and when defining the cut-off the course of illness should be taken into account.

## References

- 1 Nes, S. I. v. *Improving and standardising assessment of patients with immune-mediated neuropathies: Peripheral Neuropathy outcome measures Standardisation study (PeriNomS study – part 1)* PhD thesis, Erasmus universiteit Rotterdam, (2012).
- 2 Peters, M. J., van Nes, S. I., Vanhoutte, E. K. *et al.* Revised normative values for grip strength with the Jamar dynamometer. *Journal of the peripheral nervous system* : JPNS **16**, 47-50 (2011).
- 3 van Nes, S. I., Faber, C. G., Hamers, R. M. *et al.* Revising two-point discrimination assessment in normal aging and in patients with polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **79**, 832-834 (2008).
- 4 van Nes, S. I., Vanhoutte, E. K., Faber, C. G. *et al.* Improving fatigue assessment in immune-mediated neuropathies: the modified Rasch-built fatigue severity scale. *Journal of the peripheral nervous system* : JPNS **14**, 268-278 (2009).
- 5 van Nes, S. I., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 6 Merkies, I. S., van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *Journal of neurology, neurosurgery, and psychiatry* **81**, 1194-1199 (2010).
- 7 Aaronson, N. K. Quality of life: what is it? How should it be measured? *Oncology* **2**, 69-76, 64 (1988).
- 8 World Health Organization. *ICF : international classification of functioning, disability and health*. (World Health Organization, 2001).
- 9 Jaeschke, R., Singer, J., and Guyatt, G. H. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* **10**, 407-415 (1989).
- 10 Beaton, D. E., Boers, M., and Wells, G. A. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current opinion in rheumatology* **14**, 109-114 (2002).
- 11 Merkies, I. S., Schmitz, P. I., Samijn, J. P. *et al.* Fatigue in immune-mediated polyneuropathies. European Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **53**, 1648-1654 (1999).
- 12 Merkies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 13 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 14 Merkies, I. S. J. *Evaluation of scales and measurement instruments in immune-mediated polyneuropathies*. Thesis, Erasmus Medical Center, (2001).
- 15 Dyck, P. J., Davies, J. L., Litchy, W. J. *et al.* Longitudinal assessment of diabetic polyneuropathy using a composite score in the Rochester Diabetic Neuropathy Study cohort. *Neurology* **49**, 229-239 (1997).
- 16 Dyck, P. J., Karnes, J. L., O'Brien, P. C. *et al.* The Rochester Diabetic Neuropathy Study: reassessment of tests and criteria for diagnosis and staged severity. *Neurology* **42**, 1164-1170 (1992).
- 17 Dyck, P. J., Kratz, K. M., Lehman, K. A. *et al.* The Rochester Diabetic Neuropathy Study: design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. *Neurology* **41**, 799-807 (1991).
- 18 Mathiowetz, V. in *Muscle strength testing. Instrumented and non-instrumented systems* (ed L. R. Amundsen) 163-177 (Churchill Livingstone, 1990).
- 19 Medical Research Council. *Medical Research Council. Aids to the investigation of the peripheral nervous system.*, (Her Majesty's Stationary Office, 1943).
- 20 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet neurology* **7**, 136-144 (2008).
- 21 Dyck, P. J., Boes, C. J., Mulder, D. *et al.* History of standard scoring, notation, and summation of neuromuscular signs. A current survey and recommendation. *Journal of the peripheral nervous system* : JPNS **10**, 158-173 (2005).
- 22 Vanhoutte, E. K., Faber, C. G., van Nes, S. I. *et al.* Modifying the Medical Research Council grading system through Rasch analyses. *Brain : a journal of neurology* **135**, 1639-1649 (2012).

- 23 Vanhoutte, E. K., Faber, C. G., and Merkies, I. S. 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscul Disord* **23**, 924-933 (2013).
- 24 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of clinical epidemiology* **49**, 711-717 (1996).
- 25 Lauria, G., Bakkers, M., Schmitz, C. *et al.* Intraepidermal nerve fiber density at the distal leg: a worldwide normative reference study. *Journal of the peripheral nervous system : JPNS* **15**, 202-207 (2010).
- 26 Merkies, I. S., Schmitz, P. I., Van Der Meche, F. G. *et al.* Comparison between impairment and disability scales in immune-mediated polyneuropathies. *Muscle & nerve* **28**, 93-100 (2003).
- 27 Bril, V., Katzberg, H., Donofrio, P. *et al.* Electrophysiology in chronic inflammatory demyelinating polyneuropathy with IGIV. *Muscle & nerve* **39**, 448-455 (2009).
- 28 Kersten, P., Kucukdeveci, A. A., and Tennant, A. The use of the Visual Analogue Scale (VAS) in rehabilitation outcomes. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* **44**, 609-610 (2012).
- 29 Lund, I., Lundeborg, T., Sandberg, L. *et al.* Lack of interchangeability between visual analogue and verbal rating pain scales: a cross sectional description of pain etiology groups. *BMC medical research methodology* **5**, 31 (2005).
- 30 Thomee, R., Grimby, G., Wright, B. D. *et al.* Rasch analysis of Visual Analog Scale measurements before and after treatment of Patellofemoral Pain Syndrome in women. *Scandinavian journal of rehabilitation medicine* **27**, 145-151 (1995).
- 31 Met, R., Reekers, J. A., Koelemay, M. J. *et al.* The AMC linear disability score (ALDS): a cross-sectional study with a new generic instrument to measure disability applied to patients with peripheral arterial disease. *Health and quality of life outcomes* **7**, 88 (2009).
- 32 Vandervelde, L., Van den Bergh, P. Y., Penta, M. *et al.* Validation of the ABILHAND questionnaire to measure manual ability in children and adults with neuromuscular disorders. *Journal of neurology, neurosurgery, and psychiatry* **81**, 506-512 (2010).
- 33 Group, T. E. EuroQol--a new facility for the measurement of health-related quality of life. *Health policy* **16**, 199-208 (1990).
- 34 Aaronson, N. K., Muller, M., Cohen, P. D. *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *Journal of clinical epidemiology* **51**, 1055-1068 (1998).
- 35 Bergner, M., Bobbitt, R. A., Carter, W. B. *et al.* The Sickness Impact Profile: development and final revision of a health status measure. *Medical care* **19**, 787-805 (1981).
- 36 Erdman, R. A., Passchier, J., Kooijman, M. *et al.* The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects. *Psychological reports* **72**, 1027-1035 (1993).
- 37 Trompenaars, F. J., Masthoff, E. D., Van Heck, G. L. *et al.* Content validity, construct validity, and reliability of the WHOQOL-Bref in a population of Dutch adult psychiatric outpatients. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* **14**, 151-160 (2005).
- 38 Vickrey, B. G., Hays, R. D., and Beckstrand, M. Development of a health-related quality of life measure for peripheral neuropathy. *Neurorehabilitation and neural repair* **14**, 93-104 (2000).
- 39 Ware Jr, J. E., Snow, K. K., M., K. *et al.* *SF-36 health survey. Manual and interpretation guide.* . (The health institute, New England Medical Center, 1997).
- 40 Law, M., Polatajko, H., Pollock, N. *et al.* Pilot testing of the Canadian Occupational Performance Measure: clinical and measurement issues. *Canadian journal of occupational therapy. Revue canadienne d'ergotherapie* **61**, 191-197 (1994).
- 41 Law, M., Baptiste, S., McColl, M. *et al.* The Canadian occupational performance measure: an outcome measure for occupational therapy. *Canadian journal of occupational therapy. Revue canadienne d'ergotherapie* **57**, 82-87 (1990).



# **Nederlandse samenvatting**

## Algemene introductie en achtergrond informatie

Dit proefschrift beschrijft de resultaten van de Perifere Neuropathie uitkomstmaten Standardisatie (PeriNomS) studie. Dit is een onderzoek naar uitkomstmaten voor immuun-gemedieerde neuropathieën, waaronder het Guillain-Barré syndroom (GBS), chronische inflammatoire demyeliniserende polyradiculoneuropathie (CIDP), polyneuropathie geassocieerd met een monoklonale gammopathie (MGUSP) en multifocale motore neuropathie (MMN). Het doel van PeriNomS studie (**Hoofdstuk 1**) is het vaststellen van een basisset uitkomstmaten voor immuun-gemedieerde polyneuropathie voor gebruik in toekomstige klinische trials en follow-up studies.

De PeriNomS studie omvat drie delen.

Het eerste deel werd uitgevoerd door dr. Sonja I. van Nes; zij verdedigde haar proefschrift in januari 2011. Daarin beschrijft zij normaalwaardenstudies, vergelijking van validiteit, betrouwbaarheid en responsiviteit van verschillende uitkomstmaten; een aantal uitkomstenmaten werd ontwikkeld met behulp van de Rasch-methode en het concept 'minimum clinically important difference' (MCID) werd geïntroduceerd.<sup>1-6</sup> Dit proefschrift omvat de resultaten van het tweede deel van de PeriNomS studie, waarin 163 patiënten (GBS: 55, CIDP: 59, MGUSP: 23, MMN: 26) werden onderzocht en gedurende een periode van één jaar gevolgd. De responsiviteit van de verschillende uitkomstmaten, dat wil zeggen het vermogen om een verbetering of verslechtering te meten, werd vergeleken.

Het derde deel van de PeriNomS studie wordt momenteel uitgevoerd en wordt als onderdeel van toekomstperspectieven besproken in de discussie.

**Hoofdstuk 1** van dit proefschrift beschrijft de klinische kenmerken van de immuun-gemedieerde polyneuropathieën (GBS, CIDP, MGUSP en MMN) en de onderzoeksopzet van de PeriNomS studie. Voor zowel het cross-sectionele als het longitudinale deel van het onderzoek werden verschillende uitkomstenmaten op verschillende meetniveaus onderzocht, gebaseerd op de internationale WHO-classificatie en het kwaliteit van levenconcept (stoornissen, beperkingen, participatieproblemen en kwaliteit van leven). Ook werden de meningen van diverse experts hierin betrokken.<sup>7,8</sup>

Ondanks de bekende gebreken zijn de meeste uitkomstmaten die gebruikt worden bij onderzoek naar inflammatoire neuropathieën ordinaal en gebaseerd op de klassieke testtheorie. Het Rasch-model ondervangt de nadelen van ordinale uitkomstmaten (**Hoofdstuk 2**). Deze theorie is gebaseerd op een logische aanname: patiënten met een grotere handelingsvaardigheid hebben een grotere kans om een taak correct uit te kunnen voeren dan patiënten met meer beperkingen. Dit hoofdstuk beschrijft de verschillende voorwaarden waaraan de constructie en evaluatie van een uitkomstmaat gebaseerd op het Rasch model moet voldoen, zoals een goede statistische waarde

van items en personen, geordende antwoordopties, geen item-bias of items die aan elkaar gerelateerd zijn en unidimensionaliteit. Een uitkomstmaat op het intervallevel wordt verkregen wanneer aan alle Rasch-voorwaarden wordt voldaan.

Wetenschappelijke studies die behandelopties vergelijken, maken vaak gebruik van de aan- of afwezigheid van een statistisch significant verschil (p-waarde) tussen de groepen, bijvoorbeeld behandelgroep versus placebogroep. De aanwezigheid van een statistisch significant verschil betekent echter niet automatisch dat de bevindingen ook daadwerkelijk klinisch relevant zijn. **Hoofdstuk 3** gaat nader in op het begrip 'minimum clinically important difference' (MCID). Dit is gedefinieerd als de minimale verandering in score die noodzakelijk is om een klinisch relevant verschil te detecteren.<sup>9</sup> Er zijn verschillende methoden beschikbaar om de MCID te berekenen en er is nog geen internationale consensus over welke techniek het beste is.<sup>10</sup> De grootste beperking van de MCID-berekeningen is waarschijnlijk de aanname dat de standaard error (SE) 'statistisch' (onveranderlijk) is over het bereik van de gebruikte uitkomstenmaat. Het MCID-SE-begrip maakt gebruik van Rasch-getransformeerde data. Het blijkt dat de SE varieert over het bereik van een meeteenheid: de SE is het laagst in het midden en neemt toe aan de uiteinden van het bereik. Deze methode werd gebruikt om aan te tonen dat daarmee ook het aantonen van respondent zijn een dynamisch patroon heeft. Een respondent wordt gedefinieerd als een patiënt bij wie een significante verandering groter dan een MCID-SE score van 1.96 wordt vastgesteld.

## Studies op het WHO-niveau Stoornissen

Stoornissen worden op basis van de WHO-classificatie gedefinieerd als een of meer problemen in lichaamsfunctie of fysieke problemen structuur die kunnen bijdragen aan beperkingen in activiteit en participatie.<sup>8</sup>

Vermoeidheid wordt gezien als een invaliderende klacht bij patiënten met immuun-gemedieerde neuropathieën.<sup>11</sup> De 9-item 'fatigue severity scale' (FSS) voldoet aan wetenschappelijke eisen en wordt gebruikt om vermoeidheid te meten. Deze uitkomstenmaat is echter gebaseerd op de klassieke testtheorie en is dus een ordinale schaal. Het is dan ook niet verrassend dat de originele 9-item FSS niet aan de criteria van het Rasch-model voldeed (**Hoofdstuk 4**). De 9-item FSS werd geanalyseerd met het Rasch-model, waarbij gebruikgemaakt werd van data van patiënten met GBS, CIDP en MGUSP die in het kader van de INCAT-studies reeds onderzocht waren. Twee statistisch niet-passende vragen werden verwijderd. Daarnaast werden de zeven antwoordopties gemodelleerd tot vier categorieën. Hierdoor werd een 7-item Rasch-gebaseerde vermoeidheidschaal gecreëerd die aan alle Rasch-voorwaarden voldeed. Het wordt aangeraden deze schaal in toekomstige klinische trials te gebruiken wanneer men somscores wil vergelijken (bijvoorbeeld het effect tussen behandeling en placebo op vermoeidheid).



Het meten van spierkracht en sensibeleit wordt in klinische trials vaak als primaire uitkomstmaat gebruikt. De 'modified ISS' (mISS) en de 'Neuropathy Impairment Score sensory subset' (NISs) worden frequent gebruikt als uitkomstmaten voor het meten van sensibeleit. Spierkracht wordt vaak gemeten met instrumenten zoals de Martin Vigorimeter, of door het manueel testen van de spierkracht met behulp van het 'Medical Research Council' (MRC) scoringssysteem of het motorische onderdeel van de 'Neuropathy Impairment Score' (NISm).<sup>12-19</sup> De sensitiviteit van knijpkrachtmetingen met de Martin Vigorimeter werd aangetoond met behulp van data uit de ICE-trial. De ICE-trial is de grootste gerandomiseerde trial naar het effect van behandeling met intraveneus immuunglobuline bij patiënten met CIDP. De Martin Vigorimeter was in staat om eenvoudiger en op een eerder tijdstip in de behandeling klinisch betekenisvolle veranderingen te detecteren in vergelijking met de primaire ordinale uitkomstenmaten (**Hoofdstuk 5**).<sup>20</sup>

De MRC is in de dagelijkse praktijk de meest gebruikte methode om spierkracht van verschillende spiergroepen te meten.<sup>21</sup> In **hoofdstuk 6** wordt aangetoond dat artsen, onafhankelijk van hun ervaring of het type neuromusculaire aandoening dat onderzocht werd, niet in staat zijn om onderscheid te maken tussen de verschillende MRC-categorieën. Bij 86% van de onderzochte spiergroepen bij 1065 patiënten konden artsen het onderscheid tussen de verschillende MRC antwoordopties niet op een juiste manier maken.<sup>22</sup> Met de MRC-score kan daardoor geen adequaat onderscheid worden gemaakt tussen de verschillende gradaties van spierzwakte. Om deze problemen te ondervangen werd de MRC-score aangepast naar vier antwoordopties. Intuïtief lijkt dit de mogelijkheid te verkleinen om kleine functionele veranderingen bij een patiënt te detecteren. Maar, het vasthouden aan de (niet goed functionerende) zes antwoordopties geeft een vals gevoel van precisie en kan daarmee meetfouten vergroten. Dit zou kunnen leiden tot fout-positieve klinisch betekenisvolle verbeteringen. De vier antwoordopties zijn ordinaal; een betekenisvolle somscore kan alleen verkregen worden nadat de data getransformeerd zijn naar het intervallevel met behulp van de Rasch-methode. Nadat aan alle criteria van het Rasch-model was voldaan, werd de Rasch-getransformeerde MRC 12-spiergroepen somscore verkregen voor patiënten met GBS en CIDP.

Na Rasch transformatie werd de NISs vergeleken met de mISS en de NISm met de MRC-somscore (**Hoofdstuk 7**). De responsiviteit van de longitudinaal verkregen data van de getransformeerde schalen werd onderzocht, waarbij het MCID-SE-concept gebruikt werd. De responsiviteit van de sensibele schalen was laag en matig bij de motorische schalen. De waarden van de motorische schalen waren beter bij patiënten met GBS dan bij patiënten met CIDP. In het algemeen was er geen verschil tussen de verschillende sensibele schalen of tussen de motorische schalen. Bij patiënten met MGUPS was de responsiviteit, indien aanwezig, erg laag.

## Studies op het WHO-niveau Activiteit en Participatie

Volgens de World Health Organisation wordt activiteit gedefinieerd als het uitvoeren van een taak of actie door een individu; participatie wordt omschreven als de mate waarin iemand in staat is deel te nemen aan het sociale leven.<sup>8</sup> Omdat activiteit en participatie niet duidelijk afgebakend zijn, is er een uitkomstmaat nodig die activiteit/participatie meet bij alle patiënten met immuun-gemedieerde neuropathieën.

Hoewel GBS, CIDP en MGUSP verschillende aandoeningen zijn, vooral wat betreft het natuurlijk beloop, liet Rasch-analyse geen DIF zien, wat betekent dat patiënten met deze aandoeningen de verschillende items op dezelfde manier ervaren. Het was dus mogelijk om één Rasch-built Overall Disability Scale (RODS) te creëren voor patiënten met GBS, CIDP en MGUSP. Het indolente beloop van MGUSP uitte zich in een lagere responsiviteit. Patiënten met MMN ervaren dezelfde items anders en het was dus nodig om voor deze patiëntengroep een aparte RODS te creëren. Door items met elkaar te verbinden is het mogelijk om de uitkomsten van deze twee verschillende RODS te vergelijken.

Er werden twee verschillende RODS-schalen ontwikkeld die aan de voorwaarden van het Rasch-model voldeden: één voor GBS/CIDP/MGUS (inflammatory-RODS, I-RODS) (**Hoofdstuk 8**) en één voor MMN (MMN-RODS) (**Hoofdstuk 9**). Deze hoofdstukken beschrijven stapsgewijs de constructie van de schalen, waarbij de methodologische beperkingen besproken worden.

De I-RODS heeft een betere spreiding van de moeilijkheidsgraad van de items in vergelijking met de items van de veel gebruikte Overall Disability Sum Score (ODSS of INCAT-ONLS). Dit betekent dat de targeting van de I-RODS voor patiënten met verschillende niveaus van vaardigheid beter is. De I-RODS heeft aan de uiteinden een groter bereik en laat dus minder vloer- en plafondeffect zien.<sup>5</sup> Cross-cultureel onderzoek van de 24 items toonde item bias aan bij 5 van de 24 items. Dit benadrukt het belang van cross-culturele validatie voordat een uitkomstmaat gebruikt kan worden in verschillende geografische regio's. Ondanks deze bevindingen laat het longitudinale internationale deel van de PeriNomS studie zien dat de I-RODS in staat is om vaker klinisch betekenisvolle verandering in de tijd te meten, met een groter verschil in verandering, dan de INCAT-ONLS disability schaal bij patiënten met GBS en CIDP. De responsiviteit van de MMN-RODS was relatief laag (**Hoofdstuk 9**). Doordat IgM-MGUSP een indolent beloop heeft, was er een gebrek aan responsiviteit en kon geen verschil tussen de schalen worden aangetoond (**Hoofdstuk 10**).

Het MCID-SE-concept met Rasch-getransformeerde data liet een dynamisch patroon zien van 'respondenten' onder patiënten met een immuun-gemedieerde neuropathie. De responsiviteit van de I-RODS bij patiënten met een immuun-gemedieerde neuropathie was beter in vergelijking met andere uitkomstmaten. Er moet rekening gehouden worden met het natuurlijk beloop van een ziekte wanneer een afkappunt voor klinisch betekenisvolle verandering wordt vastgesteld. Bij aandoeningen met

een langzaam natuurlijk beloop, zoals MGUSP, kan een langere follow-upperiode noodzakelijk zijn om zo uitkomstmaten te ontwikkelen die sensitief zijn voor het meten van veranderingen (**Hoofdstuk 11**).

## De basisset uitkomstenmaten voor toekomstige studies in inflammatoire neuropathieën (Hoofdstuk 12)

Vertegenwoordigers van de deelnemende centra aan de PeriNomS studie kwamen in 2013 gedurende drie dagen bijeen in de 196th ENMC Internationale Workshop 'Outcome Measures in inflammatory peripheral Neuropathies'. Zij verkregen consensus over de specifieke basisset van uitkomstmaten die in toekomstige trials in immuun-gemedieerde neuropathieën gebruikt zal worden. Deze basisset is weergegeven in Tabel 1.<sup>23</sup>

De primaire uitkomstmaat bij patiënten met GBS, CIDP en MMN moet zich richten op het niveau van activiteit en participatie, zoals gemeten met de RODS. Bij MGUSP-patiënten werden nauwelijks veranderingen waargenomen, waardoor een definitieve basisset voor deze groep nog niet gedefinieerd kan worden. Een langere periode van follow-up waarbij meer data van een grotere patiëntengroep verkregen wordt is nodig om de basisset voor MGUSP vast te stellen.

**Tabel 1.** Overzicht van de minimale basisset en aanbevelingen

	GBS	CIDP	MMN	MGUSP
<b>Minimale basisset</b>				
Stoonissen-niveau	Martin Vigorimeter RT-mISS Ventilatie noodzakelijk (J/N) Duur van ventilatie	Martin Vigorimeter RT-mISS 'Manuele spier testen'	Martin Vigorimeter Patient-specifieke spier testen RT-MRC scores	Nog niet vastgesteld; aanvullende studies noodzakelijk
Activiteit- en participatie-niveau	<b>I-RODS*</b> GBS disability scale	<b>I-RODS*</b> Originele INCAT disability score	<b>MMN-RODS*</b>	Zie hierboven
Kwaliteit van leven-niveau		5-PGIC SF-36	RT-QoL scale	Zie hierboven
<b>Aanbevelingen</b>				
Stoonissen –niveau	RT-MRCss Originele MRCss RT-FSS	11-PI-NRS RT-FSS	-	RT-mISS
Activiteit- en participatie-niveau		-	-	RODS Originele INCAT 10-point
Kwaliteit van leven-niveau		-	-	PGIC SF36 or Euro-QoL

**Legenda Tabel 1:** 5-PGIC = 5-points patient global impression of change; 11-PI-NRS = 11-point Pain-Intensity Numerical Rating Scale; FSS = Fatigue Severity Scale; INCAT= Inflammatory Neuropathy Cause and Treatment; I-RODS=inflammatory RODS; mISS = modified INCAT sensory sumscore; MRCss = Medical Research Council sum score; RODS = Rasch-built overall disability scale; RT = Rasch getransformeerd; QoL = Quality of Life;

\*Geeft primaire uitkomstmaat aan.

## Referenties

- 1 Nes, S. I. v. *Improving and standardising assessment of patients with immune-mediated neuropathies: Peripheral Neuropathy outcome measures Standardisation study (PeriNomS study – part 1)* PhD thesis, Erasmus universiteit Rotterdam, (2012).
- 2 Peters, M. J., Van Nes, S. I., Vanhoutte, E. K. *et al.* Revised normative values for grip strength with the Jamar dynamometer. *Journal of the peripheral nervous system : JPNS* **16**, 47-50 (2011).
- 3 Nes, S. I. v., Faber, C. G., Hamers, R. M. *et al.* Revising two-point discrimination assessment in normal aging and in patients with polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry* **79**, 832-834 (2008).
- 4 Nes, S. I. v., Vanhoutte, E. K., Faber, C. G. *et al.* Improving fatigue assessment in immune-mediated neuropathies: the modified Rasch-built fatigue severity scale. *Journal of the peripheral nervous system : JPNS* **14**, 268-278 (2009).
- 5 Nes, S. I. v., Vanhoutte, E. K., van Doorn, P. A. *et al.* Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. *Neurology* **76**, 337-345 (2011).
- 6 Merkies, I. S., Van Nes, S. I., Hanna, K. *et al.* Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *Journal of neurology, neurosurgery, and psychiatry* **81**, 1194-1199 (2010).
- 7 Aaronson, N. K. Quality of life: what is it? How should it be measured? *Oncology* **2**, 69-76, 64 (1988).
- 8 World Health Organization. *ICF : international classification of functioning, disability and health*. (World Health Organization, 2001).
- 9 Jaeschke, R., Singer, J., and Guyatt, G. H. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* **10**, 407-415 (1989).
- 10 Beaton, D. E., Boers, M., and Wells, G. A. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Current opinion in rheumatology* **14**, 109-114 (2002).
- 11 Merkies, I. S., Schmitz, P. I., Samijn, J. P. *et al.* Fatigue in immune-mediated polyneuropathies. European Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **53**, 1648-1654 (1999).
- 12 Merkies, I. S. and Lauria, G. 131st ENMC international workshop: selection of outcome measures for peripheral neuropathy clinical trials 10-12 December 2004, Naarden, The Netherlands. *Neuromuscul Disord* **16**, 149-156 (2006).
- 13 Merkies, I. S., Schmitz, P. I., van der Meche, F. G. *et al.* Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology* **54**, 943-949 (2000).
- 14 Merkies, I. S. J. *Evaluation of scales and measurement instruments in immune-mediated polyneuropathies*. Thesis, Erasmus Medical Center, (2001).
- 15 Dyck, P. J., Davies, J. L., Litchy, W. J. *et al.* Longitudinal assessment of diabetic polyneuropathy using a composite score in the Rochester Diabetic Neuropathy Study cohort. *Neurology* **49**, 229-239 (1997).
- 16 Dyck, P. J., Karnes, J. L., O'Brien, P. C. *et al.* The Rochester Diabetic Neuropathy Study: reassessment of tests and criteria for diagnosis and staged severity. *Neurology* **42**, 1164-1170 (1992).
- 17 Dyck, P. J., Kratz, K. M., Lehman, K. A. *et al.* The Rochester Diabetic Neuropathy Study: design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. *Neurology* **41**, 799-807 (1991).
- 18 Mathiowetz, V. in *Muscle strength testing. Instrumented and non-instrumented systems* (ed L. R. Amundsen) 163-177 (Churchill Livingstone, 1990).
- 19 Medical Research Council. *Medical Research Council. Aids to the investigation of the peripheral nervous system.*, (Her Majesty's Stationary Office, 1943).
- 20 Hughes, R. A., Donofrio, P., Bril, V. *et al.* Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial. *Lancet neurology* **7**, 136-144 (2008).
- 21 Dyck, P. J., Boes, C. J., Mulder, D. *et al.* History of standard scoring, notation, and summation of neuromuscular signs. A current survey and recommendation. *Journal of the peripheral nervous system : JPNS* **10**, 158-173 (2005).
- 22 Vanhoutte, E. K., Faber, C. G., van Nes, S. I. *et al.* Modifying the Medical Research Council grading system through Rasch analyses. *Brain : a journal of neurology* **135**, 1639-1649 (2012).

- 23 Vanhoutte, E. K., Faber, C. G., and Merkies, I. S. 196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands. *Neuromuscul Disord* **23**, 924-933 (2013).
- 24 Stucki, G., Daltroy, L., Katz, J. N. *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *Journal of clinical epidemiology* **49**, 711-717 (1996).
- 25 Lauria, G., Bakkers, M., Schmitz, C. *et al.* Intraepidermal nerve fiber density at the distal leg: a worldwide normative reference study. *Journal of the peripheral nervous system : JPNS* **15**, 202-207 (2010).
- 26 Merkies, I. S., Schmitz, P. I., Van Der Meche, F. G. *et al.* Comparison between impairment and disability scales in immune-mediated polyneuropathies. *Muscle & nerve* **28**, 93-100 (2003).
- 27 Bril, V., Katzberg, H., Donofrio, P. *et al.* Electrophysiology in chronic inflammatory demyelinating polyneuropathy with IGIV. *Muscle & nerve* **39**, 448-455 (2009).
- 28 Kersten, P., Kucukdeveci, A. A., and Tennant, A. The use of the Visual Analogue Scale (VAS) in rehabilitation outcomes. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* **44**, 609-610 (2012).
- 29 Lund, I., Lundeborg, T., Sandberg, L. *et al.* Lack of interchangeability between visual analogue and verbal rating pain scales: a cross sectional description of pain etiology groups. *BMC medical research methodology* **5**, 31 (2005).
- 30 Thomee, R., Grimby, G., Wright, B. D. *et al.* Rasch analysis of Visual Analog Scale measurements before and after treatment of Patellofemoral Pain Syndrome in women. *Scandinavian journal of rehabilitation medicine* **27**, 145-151 (1995).
- 31 Met, R., Reekers, J. A., Koelemay, M. J. *et al.* The AMC linear disability score (ALDS): a cross-sectional study with a new generic instrument to measure disability applied to patients with peripheral arterial disease. *Health and quality of life outcomes* **7**, 88 (2009).
- 32 Vandervelde, L., Van den Bergh, P. Y., Penta, M. *et al.* Validation of the ABILHAND questionnaire to measure manual ability in children and adults with neuromuscular disorders. *Journal of neurology, neurosurgery, and psychiatry* **81**, 506-512 (2010).
- 33 Group, T. E. EuroQol--a new facility for the measurement of health-related quality of life. *Health policy* **16**, 199-208 (1990).
- 34 Aaronson, N. K., Muller, M., Cohen, P. D. *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *Journal of clinical epidemiology* **51**, 1055-1068 (1998).
- 35 Bergner, M., Bobbitt, R. A., Carter, W. B. *et al.* The Sickness Impact Profile: development and final revision of a health status measure. *Medical care* **19**, 787-805 (1981).
- 36 Erdman, R. A., Passchier, J., Kooijman, M. *et al.* The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects. *Psychological reports* **72**, 1027-1035 (1993).
- 37 Trompenaars, F. J., Masthoff, E. D., Van Heck, G. L. *et al.* Content validity, construct validity, and reliability of the WHOQOL-Bref in a population of Dutch adult psychiatric outpatients. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* **14**, 151-160 (2005).
- 38 Vickrey, B. G., Hays, R. D., and Beckstrand, M. Development of a health-related quality of life measure for peripheral neuropathy. *Neurorehabilitation and neural repair* **14**, 93-104 (2000).
- 39 Ware Jr, J. E., Snow, K. K., M., K. *et al.* *SF-36 health survey. Manual and interpretation guide.* (The health institute, New England Medical Center, 1997).
- 40 Law, M., Polatajko, H., Pollock, N. *et al.* Pilot testing of the Canadian Occupational Performance Measure: clinical and measurement issues. *Canadian journal of occupational therapy. Revue canadienne d'ergotherapie* **61**, 191-197 (1994).
- 41 Law, M., Baptiste, S., McColl, M. *et al.* The Canadian occupational performance measure: an outcome measure for occupational therapy. *Canadian journal of occupational therapy. Revue canadienne d'ergotherapie* **57**, 82-87 (1990).

# Dankwoord

De PeriNomS studie werd mede mogelijk doordat een grote groep mensen over de hele wereld de handen ineen sloeg. Dit proefschrift is dan ook tot stand gekomen met de hulp en steun van velen. Mijn dank aan ieder van hen is groot. Een aantal mensen wil ik hier in het bijzonder noemen.

Allereerst alle patiënten die aan dit onderzoek hebben deelgenomen. Keer op keer werden jullie gevraagd hetzelfde enorme pakket vragenlijsten in te vullen. Voor velen was het invullen en naar het ziekenhuis komen voor de onderzoeken een grote opgave. Dank voor jullie volharding, zonder jullie was ik niet in staat geweest de bestaande uitkomstmaten te verbeteren en nieuwe te creëren!

De promotoren: prof. dr. C.G. Faber en prof. dr. P.A. van Doorn. Beste Pieter, ik wil je hartelijk bedanken voor je betrokkenheid bij het onderzoek.

Beste Karin, jij bezit de gave om alles positief te bekijken en daarmee iedereen om je heen te motiveren. Als ik ooit bedenkelijk keek, was jouw reactie: maar Els, dat is toch leuk! en daarin had je altijd gelijk. Bedankt voor al je inspanning en positieve energie.

De copromotor: dr. I.S.J. Merkies. Beste Ingemar, jij was het meest intensief bij mijn promotie betrokken. Ik heb onze discussies (op wetenschappelijk vlak) altijd enorm gewaardeerd. Naast deze wetenschappelijke discussies was jij ook altijd in mij als persoon geïnteresseerd en ik zal onze filosofische gesprekken zeker missen. Hoewel we niet op elk moment op één lijn hebben gezeten, wist jij mij te stimuleren en het beste uit mij te halen. Jouw toewijding aan de PeriNomS studie en mijn promotie is groot. Mijn dank is minstens zo groot!

Mijn paranimfen en co-paranimfje: Marjolein, wij zijn begonnen als kamergenootjes op de onderzoekskamer van de neurologie en dat groeide uit tot een fijne vriendschap. Jij staat al ruim 5 jaar altijd voor mij klaar, of het nu gaat om een snelle kop koffie, een uitgebreide lunch, tips over het klussen in huis of praktische adviezen over promoveren. Als ik even niet meer weet hoe ik het verder moet aanpakken, weet jij met een lach altijd raad. Janneke, hoe wij elkaar precies tegen het lijf gelopen zijn kan ik me eigenlijk niet meer zo goed herinneren, ineens was je er, uit het Noorden (vanuit Maastricht gezien is alles boven de rivieren het Noorden). Met jou deel ik een aantal vreemde obsessies. Een voorbeeld: jij begrijpt dat als je je to-do lijst met een blauwe pen begint, je deze echt niet met een zwarte pen kunt afmaken... Jij bent mijn enige MSN-vriend die ook nog af en toe online komt, we hadden altijd erg veel lol tijdens onze online chatgesprekken. Caroline, bij jou kan ik mijn verhaal altijd kwijt. Jouw enthousiasme

en vrolijkheid zijn niet met woorden uit te drukken. Jij tovert altijd een lach op mijn gezicht. Onze zondagmiddagborrels waren legendarisch!

The members of the PeriNomS study, thank you very much for all your efforts including patients. Thank you for responding to my urgent (and frequent) request for sending data and completing/faxing/scanning all the forms I needed. Especially, I would like to thank the following members: prof. dr. D.R. Cornblath for your interest, support and promoting the PeriNomS study; prof. dr. K.C. Gorson, I am grateful for the many, many patients you included and your thorough and positive feedback on the papers; prof. dr. P. Van den Bergh, prof. dr. M. Lunn, and prof. dr. E. Nobile-Orazio thank you for your interest in the progress of my PhD. I especially appreciate your kind emails and support.

Naast de leden van de PeriNomS studie zijn er verschillende collega's uit Rotterdam, Amsterdam en Utrecht die een speciale bijdrage hebben geleverd aan het includeren van patiënten. Heel erg bedankt voor al jullie inspanningen!

Dr. A. Tennant, thank you very much for your advice on Rasch analyses!

Mijn voorganger, dr. S. I. van Nes, beste Sonja, jij hebt mij zeker in het begin helemaal wegwijs gemaakt in de PeriNomS studie. Ik heb heel erg veel van jou geleerd! Dank voor al je hulp en (telefonische) ondersteuning!

Mijn opvolger, Thomas Draak, beste Tim, ik wens je heel veel succes met deel 3 van de PeriNomS studie.

Het neuromusculaire Maastrichtse team: vanaf het begin was het duidelijk dat dit een hecht team is. Ik wil jullie graag allemaal bedanken voor jullie hulp en steun en een aantal mensen in het bijzonder. Beste Mieke, ik kwam bij jou als student. Bij jou kon ik ook altijd terecht over Rasch-vragen. Jij hebt de gave om complexe materie helder en begrijpelijk uit te leggen. Beste Mayienne, niet alleen heb jij mij geholpen met de inclusie van patiënten, je hebt me ook alles geleerd over het afnemen, kleuren en tellen van huidbiopten. Als er een probleem was kon ik altijd op jou rekenen. Ik waardeer jouw inspanningen enorm! Beste Brigitte, dank voor jouw vriendschap en lieve woorden! Jij staat altijd voor me klaar!

Dr. Y. Arens en prof. dr. C. De Die, opleider en vervangend opleider van de afdeling klinische genetica: beste Yvonne en Christine, bedankt voor jullie steun en begrip, dankzij jullie was het mogelijk mijn proefschrift tijdens mijn opleiding verder af te ronden.

Prof. dr. Marc de Baets: beste Marc, jij maakte een fijne werkplek voor mij mogelijk. Ook al werk jij niet dagelijks met uitkomstmaten, je toonde altijd veel interesse in mijn vorderingen en bevindingen.

De afdeling pathologie, in het bijzonder Benoit Frere, Ilse Driesman en Aline Kosten, bedankt voor jullie hulp met het kleuren van de 'huidjes'.

Kris Sieradzan, jij bent mijn grote databaseheld. Dankzij jou zijn alle vragenlijsten netjes ingescand en digitaal beschikbaar. Heel erg bedankt!

Annette Jonkhoff en Jos Hendrix wil ik graag bedanken voor hun inspanningen voor de layout van mijn proefschrift.

Natuurlijk wil ik bijzonder graag mijn familie en vrienden bedanken: ook al had ik niet altijd even veel tijd om overal helemaal bij te zijn, jullie toonden altijd interesse in mijn proefschrift. Bedankt voor jullie talloze pogingen om te proberen te begrijpen waar mijn proefschrift ook weer precies over gaat. En ja, 'ut muulijk boekske' is nu helemaal klaar! Bedankt voor jullie steun, begrip en belangstelling!

Dear Cain, yes, it is finally done... Thank you for your love, motivation and support, you are my rock!





# Curriculum Vitae

Els Vanhoutte was born on the 6th of Januari 1985 in Leuven, Belgium. She finished secondary school at the Rythovius College in Eersel in 2003. The same year, she started her medical training at the faculty of Medicine at the Maastricht University. After obtaining her medical doctor degree in 2009, she worked as a researcher on multiple sclerosis and autonomic dysfunction and as a PhD student on the outcome measures in peripheral neuropathies at the department of neurology of the Maastricht University Medical Centre.

From 2011 until 2012 Vanhoutte worked as a resident at the department of neurology, first at the Spaarne Hospital in Hoofddorp and hereafter at the Maastricht University Medical Center.

In Oktober 2012 she started working as a resident Clinical Genetics at the Maastricht University Medical Centre and started her training in December 2012 which she expects to finish in 2016.

Els Vanhoutte werd op 6 januari 1985 geboren in Leuven, België. In 2003 behaalde zij haar VWO/Gymnasiumdiploma aan het Rythovius College te Eersel. Dat jaar ging ze geneeskunde studeren aan de Faculteit der Geneeskunde van de Universiteit Maastricht. Nadat zij haar in 2009 haar artsenbul had behaald, werkte zij als onderzoeker op multipele sclerose en autonome dysfunctie en als PhD-student op uitkomstmaten bij perifere neuropathie op de afdeling Neurologie van het Maastricht Universitair Medisch Centrum.

Van 2011 tot 2012 werkte Vanhoutte als arts-assistent neurologie, eerst in het Spaarne ziekenhuis in Hoofddorp en hierna in het Maastricht Universitair Medisch Centrum.

In oktober 2012 begon zij als arts-assistent klinische genetica in het Maastrichts Universitair Medisch Centrum en in december 2012 startte zij haar opleiding tot klinisch geneticus, die zij in 2016 verwacht te voltooien.

# List of publications

## **Optimizing temperature threshold testing in small fiber neuropathy.**

Bakkers M, Faber CG, Reulen JP, Hoeijmakers JG, **Vanhoutte EK**, Merkies IS. Muscle Nerve. 2014 Oct 7

## **196th ENMC international workshop: Outcome measures in inflammatory peripheral neuropathies 8-10 February 2013, Naarden, The Netherlands.**

**Vanhoutte EK**, Faber CG, Merkies IS; PeriNomS study group. Neuromuscul Disord. 2013 Nov;23(11):924-33.

## **Rasch-built Overall Disability Scale for patients with chemotherapy-induced peripheral neuropathy (CIPN-R-ODS).**

Binda D, **Vanhoutte EK**, Cavaletti G, Cornblath DR, Postma TJ, Frigeni B, Alberti P, Bruna J, Velasco R, Argyriou AA, Kalofonos HP, Psimaras D, Ricard D, Pace A, Galiè E, Briani C, Dalla Torre C, Lalisang RI, Boogerd W, Brandsma D, Koeppen S, Hense J, Storey D, Kerrigan S, Schenone A, Fabbri S, Rossi E, Valsecchi MG, Faber CG, Merkies IS; CI-PeriNomS study group, Galimberti S, Lanzani F, Mattavelli L, Piatti ML, Bidoli P, Cazzaniga M, Cortinovis D, Lucchetta M, Campagnolo M, Bakkers M, Brouwer B, Boogerd W, Grant R, Reni L, Piras B, Pessino A, Padua L, Granata G, Leandri M, Ghignotti I, Plasmati R, Pastorelli F, Heimans JJ, Eurelings M, Meijer RJ, Grisold W, Lindeck Pozza E, Mazzeo A, Toscano A, Russo M, Tomasello C, Altavilla G, Penas Prado M, Dominguez Gonzalez C, Dorsey SG. Eur J Cancer. 2013 Sep;49(13):2910-8.

## **Vigorimeter grip strength in CIDP: a responsive tool that rapidly measures the effect of IVIG--the ICE study.**

**Vanhoutte EK**, Latov N, Deng C, Hanna K, Hughes RA, Bril V, Dalakas MC, Donofrio P, van Doorn PA, Hartung HP, Merkies IS. Eur J Neurol. 2013 May;20(5):748-55.

## **MRC sum-score in the ICU: good reliability does not necessarily reflect "true reliability".**

**Vanhoutte EK**, Faber CG, Merkies IS; PeriNomS Study Group. Muscle Nerve. 2012 May;45(5):767-8; author reply 768-9.

## **Modifying the Medical Research Council grading system through Rasch analyses.**

**Vanhoutte EK**, Faber CG, van Nes SI, Jacobs BC, van Doorn PA, van Koningsveld R, Cornblath DR, van der Kooi AJ, Cats EA, van den Berg LH, Notermans NC, van der Pol WL, Hermans MC, van der Beek NA, Gorson KC, Eurelings M, Engelsman J, Boot H, Meijer RJ, Lauria G, Tennant A, Merkies IS; PeriNomS Study Group. Brain. 2012 May;135(Pt 5):1639-49.

## **Gain of function Nav1.7 mutations in idiopathic small fiber neuropathy.**

Faber CG, Hoeijmakers JG, Ahn HS, Cheng X, Han C, Choi JS, Estacion M, Lauria G, **Vanhoutte EK**, Gerrits MM, Dib-Hajj S, Drenth JP, Waxman SG, Merkies IS. Ann Neurol. 2012 Jan;71(1):26-39.

**Revised normative values for grip strength with the Jamar dynamometer.**

Peters MJ, van Nes SI, **Vanhoutte EK**, Bakkers M, van Doorn PA, Merkies IS, Faber CG; PeriNomS Study group. J Peripher Nerv Syst. 2011 Mar;16(1):47-50

**Peripheral neuropathy in myotonic dystrophy type 1.**

Hermans MC, Faber CG, **Vanhoutte EK**, Bakkers M, De Baets MH, de Die-Smulders CE, Merkies IS. J Peripher Nerv Syst. 2011 Mar;16(1):24-9.

**Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies.**

van Nes SI, **Vanhoutte EK**, van Doorn PA, Hermans M, Bakkers M, Kuitwaard K, Faber CG, Merkies IS. Neurology. 2011 Jan 25;76(4):337-45.

**Statistische significantie of klinische relevantie? [Statistical significance or clinical relevance?].**

**Vanhoutte EK**, Faber CG, Merkies IS. Ned Tijdschr Geneesk. 2010;154:A2516.

**Improving fatigue assessment in immune-mediated neuropathies: the modified Rasch-built fatigue severity scale.**

van Nes SI, **Vanhoutte EK**, Faber CG, Garssen M, van Doorn PA, Merkies IS; PeriNomS Study Group. J Peripher Nerv Syst. 2009 Dec;14(4):268-78.



# Valorisation-addendum

Valorisation in scientific research describes the value added to the society. Besides education and scientific research, valorisation is one of the key tasks of universities. As such, it is important to explicitly state the added value of the research performed.

## 1 Social and economic relevance?

Clinical trials in patients with immune-mediated neuropathies like Guillain-Barré syndrome (GBS), chronic inflammatory demyelinating polyradiculoneuropathy (CIDP), and monoclonal gammopathy of undetermined significance related neuropathy (MGUSP) often used different outcome measures to present their results. In the past, multiple types of often ordinal outcome measures have been used in these disorders leading to different results of studies. The design and validation process of creating an outcome measure is a time consuming process with a burden to patients (they have to complete the same outcome measure several times) and with considerable costs. Furthermore, in order to obtain comparability between trials, researchers should strive to use the same existing outcome measures.

Proper attention should be given to the choice of the best outcome measure, particularly choosing outcome measures that have been tested in terms of their clinimetric properties in the disease of interest. The use of insensitive outcome measures might be the reason for trials in the inflammatory neuropathy field being negative. This could mean that the Food and Drug Administration (FDA), due to the use of improper outcome measures, is not approving treatments that might have an effect, which may hurt patients.

Most outcome measures used thus far in inflammatory neuropathies are at the ordinal level and based on the classical test theory. The Rasch model overcomes the disadvantages of ordinal-based outcome measures. We constructed new outcome measures and evaluation new and existing outcome measures using the Rasch method. These outcome measures could overcome the previously presented problems.

In addition, trial results evaluating treatment options are often driven by the presence or absence of having a significant (p-value) difference between the various (e.g. treated versus placebo) groups. However, the presence of a statistical significant difference does not always mean that the findings are clinically relevant. The minimum clinically important difference (MCID) was introduced in the current thesis as the minimum change in score necessary to reflect a clinically relevant change.[1] There are several methods available to calculate the MCID, thus showing its many faces and the lack of international consensus on this matter.[2] The biggest disadvantages of applying these MCID techniques is perhaps the standard error (SE) being considered “static (unchanged)”

throughout the range of the outcome measure being used. The MCID-SE concept using Rasch-transformed demonstrates a dynamic pattern of ‘being-a-responder’ in patients with immune-mediated neuropathies. SE values are the lowest in the middle part of the metric and increases towards the edges of the metric’s range. A responder is defined if a patient showed a significant change exceeding  $\text{MCID-SE} \geq 1.96$ .

Using the concept of MCID and responsiveness at the individual level could give physicians the opportunity to determine scientifically and objectively whether the patient is a responder or not. When the patient is not responding, one may consider to discontinue the treatment or to change to different treatment options. This would contribute to personalized medicine and could, at the same time, lower the costs in health care.

## 2 Target groups

Outcome measures are usually developed using the Classical Test Theory (CTT). However, outcome measures based on CTT may constitute items that are arbitrarily collected with ordinal response options (e.g. 0 never, 1 seldom, 2 quite often, 3 very often, 4 always).

Although the distance between the different response options appear linear (score 1,2,3,4) is the true distance between the different categories not known and most probably unequal. Physicians often consider as an example a 1-point response change for an item (e.g., from 0 to 1) equivalent to a 1-point change from 2 to 3.

Also, patients are requested to complete all items, even though some may be irrelevant or inappropriate for their level of ability. A sumscore of the scale’s items is often calculated and the obtained data generally treated as if they were linear; frequently being exposed to parametric analyses. Creating a sum of the item scores also assumes equal relevance (“weight”) of each item, which is highly unlikely. Based on these shortcomings, CTT-based outcome measures may limit the comparison of patients and study results.

Considering the shortcomings of the CTT, it is clear that a modern scientific approach is needed for the evaluation and construction of outcome measures to improve the findings in interventional trials. Using interval measures instead of ordinal scores would give a true reflection of disease impact, of differences between individuals and groups, and of treatment effects. One of the widely used approaches is the Rasch method, which was introduced by the Danish mathematician Georg Rasch. The Rasch model states that the probability of a patient being able to “correctly answer or complete” an item or task is a logistic function of the difficulty of the task and the ability of the patient to accomplish it. Rasch analysis transforms obtained ordinal scores into interval measures and places both items and patients’ parameter estimates on the same logodds units (logit) scale. Therefore a less affected patient (a patient with a higher ability) will have a greater chance to complete a more difficult item when compared

to a patient that is more disabled. In order to create an interval outcome measure using the Rasch model, several criteria have to be fulfilled. RUMM2030 is a computer program that 'checks' whether the raw data fulfil all these criteria. There are several ways to deal with items deviating from the models' expectations.

We have used a modern scientific approach, based on the Rasch model, leading to disease-specific, true interval measurement, and therefore enable highly improved evaluation and construction of outcome measures in interventional trials. There are several pharmaceutical companies, like Baxter, Talecris, Octapharma that are involved in development of immunomodulating therapy. Although creating outcome measures using the Rasch model is not novel, the outcome measures developed by our group are unique and novel. The advantages are that we provide an interval outcome measure for specific diseases. Moreover, using these outcome measures provides the opportunity to compare the results of different research or trials on the same linear ruler, thereby improving comparability of different studies.

Muscle strength testing is frequently used in clinical trials as a primary outcome measure at the impairment level and can be measured using a tool like the Marin Vigorimeter [3] or by manual muscle testing (Medical Research Council (MRC) grades.[4] The MRC is the most widely applied tool to manually measure strength of muscle groups at bedside.[5] We demonstrated that physicians, independent of their experience or the type of neuromuscular illness examined, are unable to apply the MRC grades in a proper manner.[6] Also, the MRC grading system failed to differentiate various degrees of muscle weakness. In order to solve these problems, the MRC grades were modified to a homogenous four category response options. The four response categories are still ordinal based; therefore meaningful sum scores can only be made after transforming the data through Rasch. After Rasch modelling, we were able to present a transformed modified MRC 12 muscle groups summed score for GBS and CIDP. The Vigorimeter demonstrated the ability to capture clinically meaningful changes quite early in the interventional phase when compared with the primary ordinal-based outcome measure used.[7, 8]

Although the MRC grading system is widely used, it holds several disadvantages. The Vigorimeter overcomes these advantages, however this tool only measures grip strength. These results might trigger companies to develop devices to measure strength at the interval levels. These devices should be, easy applicable, validated and not too expensive. Such a tool could also be used in clinical practice for the follow-up of patients.

### **3 Activities and products**

Creating and (re)analyse outcome measures to develop a Rasch-based tool at the interval level is a time-consuming and labour-intensive process, that warrants specific



expertise. The industry may prefer to buy our Rasch-build outcome measures rather than developing it themselves. Through paid licensing, we can offer companies (and other academia) our developed Rasch-build outcome measures including the required analysis key. Analysing the results obtained from Rasch-build outcome measures requires expertise. It takes several years before that expertise is obtained, and therefore we are able to offer this service to companies and other universities.

#### **4 Innovation**

Existing tools for generating outcome measures are usually based on Classical Test Theory (CTT). CTT method holds several disadvantages, thereby creating an illusion of interval measurement. Therefore, CTT-based outcome measures limit the comparison of patients and study results. We have used a modern scientific approach, based on the Rasch model, leading to disease-specific, true interval measurement, and therefore highly improved evaluation and construction of outcome measures in interventional trials. Creating outcome measures using the Rasch model is not novel. However, the outcome measures developed by our group are unique and novel. The advantages are that we provide an interval outcome measure for specific diseases that are validated and responsive. Moreover, using these outcome measures provides the opportunity to compare the results of different research or trials on the same linear ruler, thereby improving comparability of different studies.

#### **5 Implementation**

We offer our outcome measures for use to companies and academia, through paid licensing. Moreover, we are able to perform the required analysis of the obtained data. These tools are intended to be used in clinical trials. For obtaining a license, please contact dr. Ingemar S.J. Merkies or prof.dr. C.G. Faber.

## Literature

1. Jaeschke, R., J. Singer, and G.H. Guyatt, *Measurement of health status. Ascertaining the minimal clinically important difference*. Control Clin Trials, 1989. **10**(4): p. 407-15.
2. Beaton, D.E., M. Boers, and G.A. Wells, *Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research*. Current opinion in rheumatology, 2002. **14**(2): p. 109-14.
3. Mathiowetz, V., *Grip and pinch strength measurements*, in *Muscle strength testing. Instrumented and non-instrumented systems*, L.R. Amundsen, Editor. 1990, Churchill Livingstone: New York. p. 163-177.
4. Medical Research Council, *Medical Research Council. Aids to the investigation of the peripheral nervous system*. 1943, London: Her Majesty's Stationary Office.
5. Dyck, P.J., et al., *History of standard scoring, notation, and summation of neuromuscular signs. A current survey and recommendation*. Journal of the peripheral nervous system : JPNS, 2005. **10**(2): p. 158-73.
6. Vanhoutte, E.K., et al., *Modifying the Medical Research Council grading system through Rasch analyses*. Brain : a journal of neurology, 2012. **135**(Pt 5): p. 1639-49.
7. Hughes, R.A., et al., *Intravenous immune globulin (10% caprylate-chromatography purified) for the treatment of chronic inflammatory demyelinating polyradiculoneuropathy (ICE study): a randomised placebo-controlled trial*. Lancet neurology, 2008. **7**(2): p. 136-44.
8. Vanhoutte, E.K., et al., *Vigormeter grip strength in CIDP: a responsive tool that rapidly measures the effect of IVIG--the ICE study*. Eur J Neurol, 2013. **20**(5): p. 748-55.

